U.PORTO

FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

OUTLIER DETECTION IN ACCOUNTING DATA

João Manuel Oliveira Machado

Dissertation

Master in Modeling, Data Analysis and Decision Support System in Optimization.

Supervised by

Professor Adelaide Maria de Sousa Figueiredo

Professor João Manuel Portela de Gama

2018

Biographical Note

João Manuel Oliveira Machado was born on May 5th, in Guimarães, Portugal. He initiates in the bachelor's in Management in 2011 at Faculdade de Ciências Sociais e Humanas in Universidade da Beira Interior (UBI), accomplish this degree in 2015 with a final grade of 13 out of 20.

In 2015, João started his professional career in Financial auditing at Cruz Martins & Associada SROC, for a period of one year.

In 2016 was admitted in the Master in Data Analytics at FEP, with the main objective of extending his knowledge of analytical techniques and tools, such as R, Python, SPSS and SQL, fundamental for his professional development.

Acknowledgements

To my supervisors, Professor Adelaide Figueiredo and Professor João Gama, for their attention and availability, for the support along this path, as well as their knowledge and guidance.

To Dra. Ana Bárbara for the granted opportunity to work in partnership with the Bank de Portugal and for her mentoring throughout the developed work, but also for presenting the theme that inspired my dissertation.

To all my friends that even far, always gave me the courage and strength to continue and finalize this stage.

To all my colleges, for all shared experiences and for all the classes and group assignments during these two years. I would like to especially thanks to my college José Medeiros and Bernardo Lemos, for sharing their knowledge and for encouraging me to extend my knowledge in programming languages.

To my parents, António and Fernanda, for their efforts to provide the best possible better and good educational background. Without you none of this would be achievable. Thank you both for the unconditional love, patience, tenderness, understanding and support. I know they will always be there for me.

To my brother, for the memories of my childhood. Not only for the pranks, games and laughs we shared, but also for the fact that it was you who shared everything.

To my girlfriend, for all the love, tenderness, patience and unconditional support at all times. Thank you for believing in me and never let me give up.

Abstract

This dissertation is the outcome of an internship at the Central Balance Sheet Data Office of Banco de Portugal (BDP). BDP publishes statistics and studies on the accounting data of non-financial corporations in Portugal. Before proceeding to its publication, the data is submitted to a quality control procedure, which has an automatic and manual component. In order to make the selection process of companies for manual analysis more efficient, several techniques of outlier detection were applied, with the purpose of optimizing the ratio between the cost of manual validation and the impact in the quality of the information. This analysis focus on a set of balance sheet attributes reported in the Quarterly Non-Financial Companies Survey (ITENF) for the period of 2010 to 2017. Firstly, statistical methods were implemented, such as the adaptation of the z-score proposed by IBM, and the boxplot with an adjustment regarding the definition of the "whiskers", based on the interdecile range for data with low variability. At the multivariate level, were implemented proximity-based methods (density-based and clustering) such as Local Outlier Factor (LOF) and DBSCAN. The results suggest that the implemented methodologies can complement each other in the selection process of companies, aiming to optimize the advantages of the different methods. At univariate level for companies with less than 20 observations the z-score should be applied and for companies with at least 20 observation the boxplot is more appropriate, because it is more robust and tends to deal better with high variability data. Nonetheless, it requires more observations so its "whiskers" can have meaning. At the multivariate level, the LOF is the most appropriated method. Although it presents more volatile results in relation to the parameterization, its output is a score, allowing for a greater flexibility in the choice of companies for manual validation.

Keywords:

Accounting data, boxplot, DBSCAN, LOF, outlier detection, quality control and z-score.

Resumo

Esta dissertação foi realizada no âmbito de um estágio curricular na Central de Balanços do Banco de Portugal (BDP). O BDP publica estatísticas e estudos sobre os dados contabilísticos das sociedades não financeiras em Portugal. Antes de proceder à publicação dos dados, os mesmos são submetidos a um procedimento de controlo de qualidade, o qual possui uma vertente automática e manual. Com o objetivo de tornar o processo de seleção de empresas para análise manual mais eficiente, serão identificadas diversas técnicas de deteção de outliers, de modo a otimizar a relação entre o custo de validação manual e o impacto na qualidade estatística. A análise efetuada irá incidir sobre um conjunto de variáveis de balanço reportadas no Inquérito Trimestral às Empresas Não Financeiras (ITENF) para o período 2010 a 2017. Ao nível univariado foram implementados métodos estatísticos, tais como a adaptação do z-score proposta pela IBM e o boxplot com uma adaptação da definição dos seus "bigodes" baseado na amplitude interdecil para dados com pouca variabilidade. Ao nível multivariado foram implementados métodos de proximidade (densidade e clustering) como o Local Outlier Factor (LOF) e o DBSCAN. Os resultados sugerem a complementaridade das várias metodologias no processo de seleção de empresas, tendo como objetivo otimizar as vantagens dos vários métodos. Ao nível univariado e para empresas com menos de 20 observações deve ser aplicado o z-score, para as restantes o boxplot é mais adequado por ser um método mais robusto e que lida melhor com a grande variabilidade nos dados. Contudo necessita de mais observações para que os seus "bigodes" tenham significado. Ao nível multivariado, o método LOF é o mais adequado. Apesar de apresentar resultados mais voláteis face à parametrização, o seu output é um "score", permitindo uma maior flexibilidade na escolha das empresas para validação manual.

Palavras-Chave:

Boxplot, dados contabilísticos, controlo de qualidade, DBSCAN, deteção de outliers, LOF e z-score.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

Over the years, there has been an increase in the capacity to store, manipulating and analyzing data. Consequently, the access to more data allows the extraction of important knowledge generally through statistical and machine learning methods, to support strategic decisions and generate competitive advantages (Witten *et al.*, 2016).

The detection of outliers has been studied by the statistics community as least since the beginning of the 19th century (Edgeworth, 1997). Around the 90's, there was an increase in interest on outlier detection and a development of new methodologies for different types of data (Kruskal, 1988).

Prior to 2000, the outlier detection task essentially involved a process of data cleaning, where anomalous observations were removed from the data (Goldstein & Uchida, 2016). However, the focus and motivation has been changing, in some areas the mechanisms that generate an outlier become the focus of the study. Therefore, an outlier has become an important source of information to the researchers, instead of being an anomaly or bad data problem (Kruskal, 1988).

In machine learning, the detection of outliers has always been of great interest (Goldstein & Uchida, 2016). Outlier detection algorithm has been extended to several application domains and often used as an improvement to traditional rule-based detection systems. This task recurs to standard algorithms, which are flexible, and can be adapted to different problems rather than following strict queries (Wit, 2016). For most outlier detection applications, the nature of the outlier is not known, making unsupervised algorithm more suitable to deal with this problem since their learning is based on the intrinsic information of the dataset (Chandola *et al.*, 2009).

In this dissertation univariate methods were applied, such as z-score and Boxplot, and multivariate methods, such as LOF and DBSCAN. The choice of methods considered the specificities of the dataset, such as the reduced number of observations per company, the fact of the dataset being composed by a set of companies with different behaviors and the availability of implementations in the R software.

The R software is a statistical open source language with an environment for statistical computing and graphics. The R software used in this dissertation was in version 3.4.3 and

can be purchased for free at CRAN (The Comprehensive R Archive Network) at http://cran.r-project.org. R is a powerful tool and object-oriented programming language, allowing the user to create functions and routines for manipulation and data analysis [1].

1.1 Motivation

The main research domain in this dissertation is the financial accounting. Most of the studies performed are related to the area of financial audit and quality control (Sharma & Panigrahi, 2012). The need to find an effective way to detect outliers in these areas has gained considerable attention from investors, academic researchers, the media, the financial community and regulators in order to prevent cases such as Enron, Lucent, WorldCom, Parmalat, YGX, SK Global, Satyam, Harris Scarface and HIH (Wong & Venkatraman, 2015).

Outlier learning in large-scale accounting data is one of the biggest challenges in the financial area. The detection of accounting outliers is a very challenging problem, since this type of data can be influenced by multiple causes, such as macroeconomics changes, accounting skullduggery and political unrest (Yuting, 2014). Currently, this task is mainly performed by rule-based detection systems, often related with previously known scenarios. Although successful, these rules often fail to generalize and adapt to different situation due follow strict queries (Schrever *et al.*, 2017).

Consequently, data mining based financial fraud detection and fraud control grant a great support since deals with large data volumes and complexities, automating processes and reducing the manual work of screening and checking various statements (Sharma & Panigrahi, 2012).

Although many studies have been carried out in the financial domain, few have resorted to the use of accounting data. In a related work, Wit (2016) perform an outlier detection in transaction level data by using a K-NN (k-Nearest Neighbor), uCLOF (unweighted Cluster Based Local Outlier Factor) and one-class SVM (Support Vector Machine). The objective was to detect transaction with extreme values, infrequent categories and combinations. Thiprungri & Vasarhelyi (2011), also perform outlier detection in transaction level data for payments. However, they recur to clustering analysis through K-means algorithm, with the objective of detect extreme values and long lags between payments and submissions. Vierdhagriswaran *et al.* (2006), perform accounting fraud detection on quarterly and annual

financial reports from Securities and Exchange Commission. These authors recur to K-means and locally weighted logistic regression.

The motivation for this dissertation is the opportunity to perform a practical implementation of several outlier detection techniques suited for financial statements on a real database, which is still a rarely addressed problem in the literature. Along with an internship at a reputable entity such as Central Balance-sheet Data Office (CBO) of BDP.

1.2 Problem Definition

The Central Balance Sheet Database (CBSD) of Banco de Portugal is an economic and financial database on Portuguese non-financial corporations. It contains information used in the compilation of statistics for economic and financial research. CBSD information is important to different stakeholders. On the one hand, CBSD provides Banco de Portugal with useful data for carrying out its tasks in terms of statistics, financial stability analysis and research on the Portuguese economy. On the other hand, CBSD gives corporations useful information for management, for example, a perception of their position within the sector of economic activity.

The data sources used to feed the CBSD are based on annual and quarterly accounting data on an individual basis. The focus of this dissertation will be on quarterly data, based in the information reported on Quarterly Survey on Non-Financial Corporation (ITENF).

The quality of the information is one of the key tasks of Banco de Portugal. Data quality is determined by factors such as accuracy, completeness, reliability and relevance. Nowadays, the higher volume and speed of arrival of new data became a greater challenge for statistical authorities like Banco de Portugal (BDP). Quality problems can lead to big economic damages and expenses.

After receiving the reports of companies, the quality control process is initiated to ensure the consistency of the accounting data in the financial year and its temporal consistency. Quality control is one of the costliest activities in the production of statistical data for economic analysis (Battipaglia *et al.*, 2004). To ensure the quality of the produced statistics BDP has 3 sequential levels of data control. The first level is done automatically at the submission of the ITENF in the online platform, e.g., variable control limits. The second level occurs also automatically when the data enters in the BDP database, and a series of statistical procedures

are implemented, e.g., temporal variation and internal coherence. It is also implemented a sequence of automatic corrections based on a cross-information with other sources of information such as INE (Statistics Portugal), companies report, ministry of justice, financial institutions, ministry of finance and IES (simplified business information). In the third level, the companies that remain with great variations or with significant differences to other sources of information will be selected for manual validation.

The quality control process is a real economic problem: the resources are limited (tight schedule and few human resources) and the needs are large (information of 4000 companies per quarter). Aware that it is not possible to manually validate all surveys, Banco de Portugal uses a set of basic algorithms based on the identification of anomalous variations and crossing data (with other statistical databases) in order to identify a set of situations that are manually validated by skilled workers.

The purpose of this work is to use a set of tools called outliers detection in the identification of situations that should be subject of manual analysis. By identifying a set of outliers, this work allows Banco de Portugal's quality control process to be improved, promoting a better allocation of scarce resources. The use of outlier's detection tools, for example multivariate outliers, allows us to identify situations that otherwise would not be detected and eliminate other situations identified by the current implemented system. At the end the process is improved, resources will be better allocated and the quality of information will be enhanced.

1.3 Dissertation structure

In terms of structure this dissertation will be divided in the following five chapters: introduction, literature review, methodology, results and conclusion. In chapter 2 will be addressed the definition of outlier and outlier detection, followed by a review on the main aspects that influence the choice of the appropriate methodology for the problem. It will be presented a review on unsupervised outlier detection methods, followed by the description of the methods used in the practical implementation. In Chapter 3 will be done the description of the dataset, some data pre-processing tasks, such as data cleaning and normalization for multivariate methods, and the presentation of the practical implementation structure for the methods described in chapter 2. In chapter 4 will be presented the results obtained for the different implementations used. A comparison will be performed between the methods, as well as an analysis of the detected outliers by period, CAE and attribute.

Chapter 5 will review the main conclusions drawn from the analysis of the results and future lines of research.

Chapter 2 - Literature Review

2.1 Outlier and Outlier detection

2.1.1 Definition of outlier

Throughout the years many outlier definitions have been presented in the literature. Its definition can be influenced by multiple factors, such as the field of study in which the analysis is being developed (statistics, machine learning, etc) or by the type or data structure that is being used. However, the most common definition used in the literature was given by Hawkins (1980):

"An observation which deviates so much from other observations as to arouse suspicious that was generated by a different mechanism".

Nevertheless, when considering the financial domain Singh & Upadhyaya (2012) define an outlier as "An anomaly is a data instance that is rare in the dataset compared to the normal instances and does not conform with a well-defined normal course of business"

In statistics and data mining, outliers are often addressed as nonconforming patterns, anomalies, extreme values, peculiarities, discordant, deviants, fraud and even noise. In a database, outliers can be generated by fraudulent cases, misreporting or recording errors, missing values, misinterpretations, natural deviations in the population and by changes or faults in the systems (Hodge *et al.*, 2004; Ghosh *et al.*, 2006).

Outlier behaviour is not synonymous of fraudulent behaviour, though it can be used as an indicator that measures the probability of this being it (Lu, 2007). Bay *et al.* (2006) in their investigation about the detection of irregularities in accounting data, stated that a significant number of the detected outliers are not the analyst interest, mainly because companies may record a business transaction in a different manner from other companies for perfectly valid reasons and companies have several events that occur only once in the scope of the collected data.

2.1.2 Definition of outlier detection

Anomaly detection is the task of identifying observations with characteristics that significantly differ from the rest of the data (Tang *et al.,* 2002). Goldstein & Uchida (2016)

suggest that two main assumptions in outlier detection are that anomalies are rare events and different from the norm of their features.

According to Davidson (2007), outlier detection has uses in multiple domains, however in most of domains the basic steps remain the same:

1. Identifying an anomaly by calculating some "signature" of the data;
2. Determine some metric to calculate an observation´s degree of deviation from the signature;
3. Set a criterion, which if exceeded by an observation´s degree of deviation makes the observation anomalous.

In the majority of outlier detection applications, the data is fitted into a model or distribution. Significant deviations from the fitted model are recognized as anomalies. However, significant deviations have a very subjective meaning since the threshold or boundary definition in the model can influence the classification of a data point as an outlier or noise. Thresholds will also determine what is accepted to be a strong or weak outlier (Aggarwal, 2017; Knorr & Ng, 1999);

There is a wide variety of applications for outlier detection related with financial analysis, such as Credit-Card fraud, insurance, bankruptcy, stock prediction, loan decision and money laundering (Ahmed *et al.*, 2016). When considering other domains, the most referred are health care (Deneshkumar *et al.,* 2014), intrusion detection (Dali *et al.*, 2015), earth science (Flach *et al.,* 2017), fault detection (Venkatasubramanian *et al.*, 2003), disastrous weather predictions (Flach *et al.,* 2017), surveillance (Diehl & Hampshire, 2002) and structural defect detection (Liu *et al.*, 2015).

2.2 Aspects of Outlier Detection Problem

As stated above there are multiple outlier detection applications, and each one has different approaches. As shown in Figure 1, some aspects like the type of input data, the label availability and constrains or requirements induced by the application domain, may influence the way the system deals with the outlier problem (Chandola *et al.*, 2009; V. Hodge *et al.*, 2004).

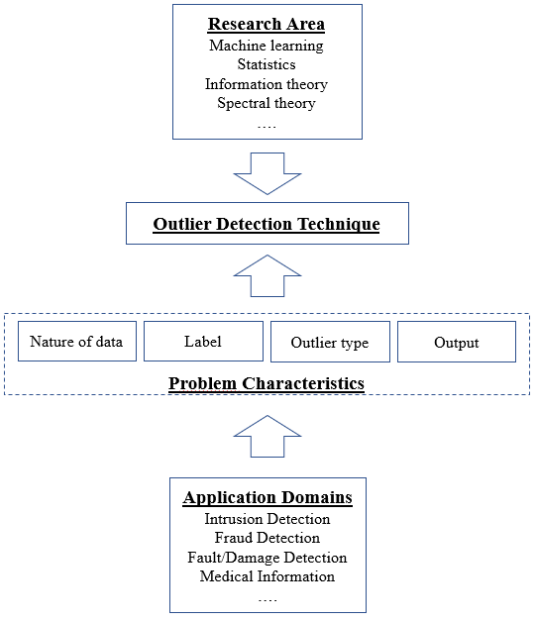

Figure 1 - Key components associated with an anomaly detection technique (source: Chandola *et al.*, 2009).

2.2.1 Input Data

The definition of outlier is directly related to the type and structure of data being analysed. Each observation can be described by a set of attributes (multivariate) or by a unique attribute (univariate) (Tan *et al.*, 2005). These attributes are mainly designed for a single data type such as continuous, binary and categorical. Due to the increase in the variety of data collected, most of the data structures are composed by heterogeneous data with mixed types of data (Do *et al.*, 2016).

The nature of data attributes determines the applicability of outlier detection techniques. For example, in statistical models for categorical data a discrete Bernoulli distribution may be used, while for continuous data a Gaussian or Gaussian mixture model can be used. Similarly, in distance-based techniques, the nature of attributes would determine the distance measure to be used (Chandola *et al.*, 2009), although in categorical data it is preferable the study of similarities (Boriah, 2008). On the other hand, the proximity-based approaches do not make any assumption about the data (Tan *et al.*, 2005).

Input data can also be categorized based on the relationship between data instances. The majority of outlier detection methods deal with point data which can be treated independently. Nevertheless, data instances can also be related temporally (time series), spatially (spatial data), or through explicit network relationships (graph data) (Aggarwal, 2017; Chandola *et al.*, 2009).

2.2.2 Type of Outlier

Finding deviations or patterns in the data that do not conform to a well-defined notion of normality is the principle of outlier detection. An important task of outlier detection techniques is to define the nature of the desired outlier. Therefore, an outlier can be classified into the following three categories (Chandola *et al.*, 2009):

Point Outlier - is the most referred type of outlier in the literature, because it occurs in most of the applications. Point outlier is an individual data instance that has an anomalous behaviour when comparing with the rest of the data. This outlier definition is widely used in statistics and often related with extreme-values (ex: when considering the age distribution of a certain population, an individual with more than 100 years should be considered a point outlier).

Contextual Outlier - Song *et al.* (2007) was the first to consider the influence of context, problem formulation and data structure, on the instance behaviour. This type of outlier is usually addressed as contextual or conditional outlier and is defined through contextual and behavioural attributes (ex: high temperatures in winter or low sales volume on the black Friday).

Collective Outlier - is a collection of instances that have an anomalous behaviour when comparing with the rest of the data. However, the instances by themselves might not be an outlier (ex: when a stock price remains the same for a long period). Usually, this type of outlier is transformed into a contextual outlier, because this type of outlier can only occur in related data instances.

2.2.3 Data Labels

The objective of labels in data instances is to classify the instances as normal or abnormal. The labels should be representative of all the type of anomalous behaviour in the dataset. Commonly, labelling data is done manually by a human specialist, which requires a huge cost of time and effort to classify each instance taking into account the different types of anomalous behaviours (Chandola *et al.*, 2009). Based on the availability of labels in the data, outlier detection techniques can operate in the three following categories illustrated in Figure 2 (Chandola *et al.*, 2009; Goldstein & Uchida, 2016):

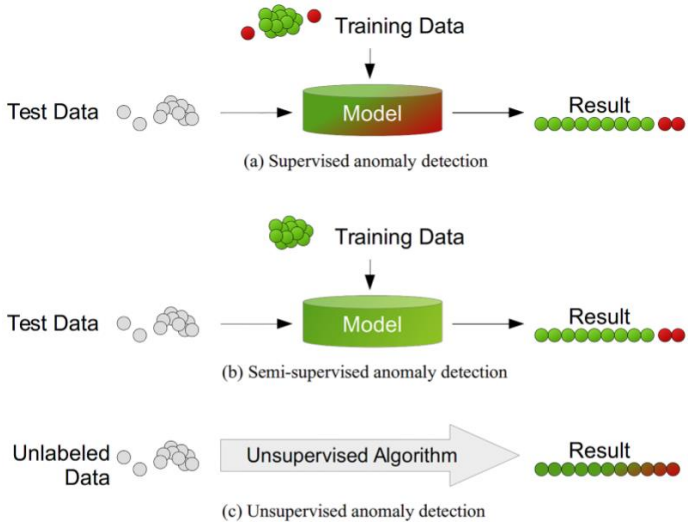

Figure 2- Different outlier detection modes depending on the availability of labels in the dataset (Source: Goldstein & Uchida, 2016).

Supervised outlier detection - assume the complete availability of labels that usually are in the binary form (i.e., normal and outlier) in the training data. The goal is to enhance a learning classification model with specific domain knowledge, to predict the class at which instance it belongs to. In theory, supervised methods should achieve superior detection rates than semi-supervised methods, since they have access to more information (Omar *et al.*, 2013). There are two major issues when performing supervised outlier detection techniques, imbalanced class distribution (Chawala *et al.*, 2004; Joshi *et al.*, 2001; Joshi *et al.*, 2002) and accurate and representative labels (Abe *et al.*, 2006). Although there are several classification algorithms not all are suitable for this task. The most used and better performing algorithms are the Support Vector Machine (SVM) (Wu & Chang, 2003; Deshpande, 2002) and Artificial Neural Network (ANN) (Sharma, 2013).

Semi-supervised outlier detection - operate with both label and unlabeled data by finding exceptional instances in the data with the use of some labeled examples. This learning approach can provide to the algorithm some supervision information while reducing the need for expensive label data (Gao *et al.*, 2006; Daneshpazhouh & Sami, 2013). In the literature semi-supervised are specially addressed by two type of data labels disposal (Aggarwal, 2017). The first is positive and unlabeled data, when the data set only contain label examples for one of the classes, positive or negative (Zhang *et al.*, 2005; Elkan *et al.*, 2008). The second occurs when the data set contains only a percentage of both classes of labels (Szummer, 2002). The most popular algorithms in this type of supervision are known as "One-class" classifiers, being the most used the one-class SVM (Khan & Madden, 2009) and Isolation Forest (IS) (Ding & Fei, 2013; Liu *et al*, 2008).

Unsupervised outlier detection – it is the most flexible setup because does not require a training set. Assumes that most of the instances on the data set perform a normal behaviour and scores the data based on the inherent dataset properties (Chandola *et al.*, 2009; Goldstein & Uchida, 2016). The main problem of unsupervised outlier detection is that usually suffers from a high false alarm rate (Xue *et al.*, 2010). This dissertation focus in unsupervised methods. In the next chapter a more extensive overview will be made on existing algorithms.

2.2.4 Output

When applying an outlier detection algorithm, it is crucial to have meaningful output for interpretation, comparison and combination. Typically, the output produced by an outlier detection algorithm is a binary label, indicating whether an instance is an outlier or not, or a score representing the "outlierness" degree of a given instance. Outlier scores can also be converted into binary labels by imposing a threshold based on the statistical distribution of the scores (Aggarwal, 2017; Gao & Tan, 2006).

2.3 Difficulties/Challenges in outlier detection

As stated previously, the outlier detection task is dependent on several aspects that may influence the approach to a problem. In this manner, when approaching the problem, it must be taken in consideration the following key challenges:

Definition of Normal instance -Distinction between normal and abnormal instances is the core of the outlier detection problem. However, it is very difficult to classify all types of normal behaviour present in a specific domain (Pahuja & Yadav, 2013). The normal behaviour present in the data may evolve over time, i.e. what is considered a normal behaviour now may not be in the future (Ahmed *et al.*, 2016). Thus, a model that suits the data previously may not be appropriate in the future. Therefore, the quality of the outlier detection method is directly dependent on the modelation of normal behaviour over time (Pahuja & Yadav, 2013).

Specific domain application - The majority of outlier detection techniques were developed to be applied in a specific domain. Although some outlier detection techniques can be adapted to other domains, there is a lack of effective techniques that can be applied in the general domain (Pahuja & Yadav, 2013; Ahmed *et al.*, 2016).

Noise Data - The presence of noise is very common in real data. Its significant presence acts as an obstacle to data analysis. Noise can be classified as abnormal behaviour in the data that is not of analyst interest (Saini *et al.,* 2016). When dealing with sparse data it is harder to identify a true anomalous deviation in the dataset. Outlier detection algorithms usually rely on a quantified measure of the outlierness of a data point or pattern (Aggarwal, 2017).

Labeled data scarcity - Outlier detection is mostly an unsupervised problem, given that in most real-world applications examples of outlier are not available (Aggarwal, 2017). Therefore, the accuracy of most techniques is evaluated using synthetic data. Despite the generation of synthetic data that could be adapted to a specific domain, both statistical and behavioural differences exist when compared with real data (Ahmed *et al.*, 2016).

Understandability - Another challenge in outlier detection is the understandability of the results, which can help define the concept of outlier and understand why these instances are outliers (Pahuja & Yadav, 2013).

Masking effect - Some outlier detection techniques are less robust to deal with mask effect. Some malicious activities may hide or adapt their anomalous behaviour by imitating or camouflage the normal behaviour (Pahuja & Yadav, 2013; Ahmed *et al.*, 2016).

Evaluation and comparison – Although outlier detection is an unsupervised task, some label information is still required to evaluate and compare the effectiveness of the methods (Goldstein & Uchida, 2016).

2.4 Outlier detection methods

Outlier detection is mostly an unsupervised task due to the lack of labels and *apriori* knowledge about the data. As shown in chapter 3, the practical implementation of this dissertation will be done in unlabeled multivariate panel data. Bramati & Croux (2007), stated that in panel data two types of outlier can be present, which are the vertical and block outlier, also known as univariate and multivariate outlier. A multivariate outlier may contain or not univariate outliers.

In terms of taxonomy, unsupervised outlier detection methods can be categorized mainly in three groups as shown in Figure 3: Nearest-Neighbour based techniques, clustering-based methods and statistical algorithms (Chandola *et al.*, 2009; Zhang *et al.*, 2007). Nonetheless,

nearest-neighbour and clustering-based techniques can be categorized as proximity-based techniques, as they define an observation as outlier based on their neighbour's proximity. Nearest-Neighbours techniques can still be divided into two categories based on different definitions of proximity: Distance-based for global proximity and Density-based for local proximity (Aggarwal, 2017). Another essential taxonomy in outlier detection is related with the *apriori* knowledge of the probability density function (pdf), parametric and non-parametric models (Markou & Sigh, 2003).

Parametric methods

These methods are modeled based on *apriori* knowledge of the data distribution (e.g., Gaussian distribution) and their parameters can be chosen based on the data means and covariance. These methods flag as outlier an observation that deviates meaningfully from the assumed data distribution (Zhang *et al.*, 2007). However, they are frequently unsuitable in high dimensionality and in real data applications as data distribution is often unknown (Markou & Sigh, 2003).

Non-Parametric methods

These methods are also known as model-free (Hodge *et al.*, 2004). They are more flexible and autonomous because no assumption is made about the statistical properties of the data. Non-Parametric methods flag outliers based on full dimensionality distance between observations (Zhang *et al.*, 2007). Thus, they usually consider the definition of a selection interval or criterion. In other words, any observation that is outside of this range or do not respect the criterion will be considered an outlier (Seo, 2002). Non-parametric tests are usually easy to use and to interpret (Laurikkala *et al.*, 2000). However, the parameters of the model are difficult to choose appropriately, and they are time and computational consuming in high dimensional data.

It should be also mentioned the existence of semi-parametric methods. Also, these methods do not make any statistical property assumptions about the data. However, they train a network model or feature space and classify as outlier an observation that deviates from the trained model. Supervised methods are related with classification techniques such as neural networks and support vector machines (Zhang *et al.*, 2007), but are outside the scope of this dissertation.

Proximity-based techniques are multivariate methods that make no assumption about the data distribution and classify observations based on different definitions of proximity. On other hand, statistical methods are generally univariate methods divided between parametric and non-parametric methods.

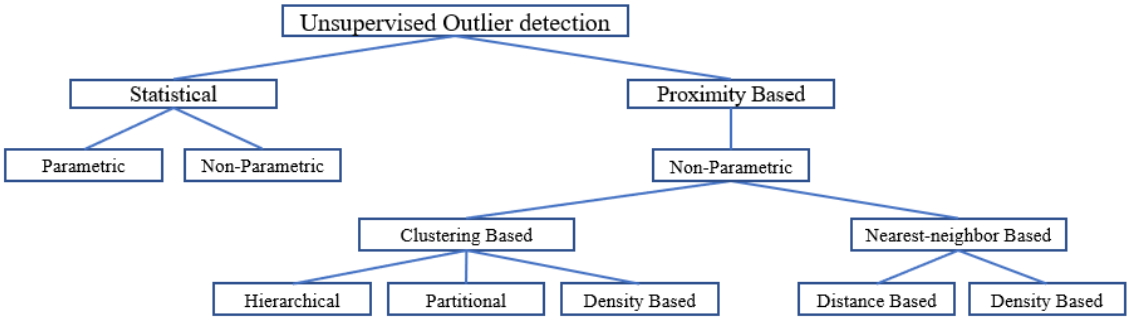

Figure 3 - A taxonomy of unsupervised outlier detection.

2.4.1 Univariate Statistical Methods

Distribution Based Methods (Parametric)

There are two types of outlier detection methods when considering a Statistical approach: formal (parametric) tests and informal (non-parametric) tests (Laurikkala *et al.,* 2000). Some of the most formal tests approached in the literature are Dixon test (Rorabacher, 1991; Iglwics and Hoaglin, 1993), Grubbs test (Grubbs, 1969), chi-squared test (Dixon, 1950), t-test, ANOVA (Kozac, 2009) and Shapiro-wilk (Iglwics & Hoaglin, 1993). This type of methods relies on hypothesis tests, also known as discordancy tests (Barnett and Lewis, 1994). Typically involves testing if whether an observation deviates from a general population which is usually represented for a well-behaved distribution such as normal, exponential or gamma (Seo, 2002). The major problem with this type of approach is that the largest real-world data has an unknown distribution or does not follow a well-behaved distribution. Other limitations are that they are susceptible for masking or swamping problems (Acuna & Rodriguez, 2004) and most of hypothesis tests are univariate methods that only detect the most extreme values (Grubbs, 1969; Dixon, 1950).

Nonetheless, within the class of parametric methods there are approaches based on mixture models. These techniques model the data based on a mixture of a set of parametric distributions (Chandola *et al.*, 2009). The Gaussian Mixture Model (GMM) is the mostly used in the literature. Normally, the parameters of the mixture models are estimated with the use

of the Expectation-Maximization (EM) algorithm, based on the maximum likelihood (ML) or maximum a posteriori (MAP) (Barkan & Averbuch, n.d.).

Regression Model-Based (Parametric)

Outlier detection using parametric regression techniques has been broadly addressed in time series data (Gupta *et al.*, 2014). The two basic steps in regression model-based outlier detection are the construction of a regressive model that properly fits the data and the comparison of each data instance against its forecasted model value (Zhang, 2013). Commonly, these methods generate a score for each instance based on its residual value. The residual value represents the part of an instance that cannot be explained by the regressive model (Chandola *et al.*, 2009).

A wide variety of regression statistical models were proposed in the literature, being the most basic the linear regression (Cook, 1977), which are not robust to the presence of outliers. Contrarily, autoregressive based models such as autoregressive moving average (ARMA) (Tiao, 1985), autoregressive integrated moving average (ARIMA) (Chen & Liu ,1993; Tsay *et al.*, 2000) and vector autoregression (VARMA) are robust as they ignore outliers when generating the fitting model and associated them with large residual values (Zhang, 2013; Rousseeuw & Leroy ,2005).

Graph Based Methods (Non-Parametric)

Graph based methods make fewer assumptions about the data, they recur to a graphical display of the data to exploit the data distribution and identify outliers as observations that are highlighted in specific positions. Regardless of being easy to implement and interpret, they suffer from a lack of precision and often visual analysis requires expert knowledge, which makes the process more resource consuming (Zhang *et al.*, 2007).

Notwithstanding, the existence of numerous graph-based outlier detection methods, the most referred and adapted in the literature are the histogram, boxplot and scatter-plot. The Histogram is the simplest method, the data is arranged into bins of equal width based on the frequency between the minimum and maximum interval values. Very small bins are reported as outliers (Aggarwal, 2017). The boxplot is probably the most popular method in outlier detection. Several adaptations of this method were made throughout the years, being the most notable the notched Boxplot, Vase plot, Bean plot and violin plot (Wickham & Stryjewski, 2011). A comprehensive explanation of this method will made further in this

chapter. The scatter-plot is a visual method that displays the spatial relation between a pairwise of attributes. As there is no defined rule to mark a point as outlier they are commonly chosen intuitively by the analyst (Aggarwal, 2017).

Kernel Function Based Methods (Non-Parametric)

The only difference between Kernel Function methods and the previous described parametric methods, is density estimated technique used (Chandola *et al.*, 2009). This method uses a kernel function to estimate the probability distribution function (pdf) of majority data instances. Instances that do not fit the new generated density function are considered potential outliers. Like other parametric methods, this method is not appropriate for a multimodal distribution, which is common in real life application (Latecki *et al.*, 2007).

Robust Estimators

When using statistical methods for outlier detection the mean and standard deviation are usual estimators of the data location and shape. Nevertheless, when the data are contaminated with outliers those estimators may significantly influence the performance of the methods used (Ben-Gal, 2005).

To measure the robustness of an estimator such as the mean or median is commonly applied the definition of the breakdown point introduced by Hampel (1971) (Iglwics & Hoaglin, 1993). The breakdown point of an estimator is the smallest proportion of data that can be changed arbitrarily without causing a noticeable impact on the statistics of interest (Finch, 2012). So, if an estimator has a breakdown point of 0.1, it means that for the estimator to be affected, the sample should be composed by more than 10% of outlier observations. That is, the larger the value of the breakdown point, the more robust is the used estimator.

2.4.1.1 Z-score

The z-score is calculated for each data point, considering the mean and standard deviation of each attribute. This method can be used for detection of extreme values, or even to normalize the data, allowing the outputs to have a comparable scale and avoiding attributes with a larger spectrum of values in multivariate methods to have a higher weight in the result (Hawkins, 1980).

The intuition behind this method is that after the data are centred and scaled, each attribute has mean 0 and standard deviation 1. Observations that are very distant to 0 should be considered an outlier. Thus, if the z-score of an observation is higher than 3, means that the

observation differs from the mean more than 3 times the standard deviation and could be an outlier (Shiffler, 1988). The z-score can be defined as:

$$Z = (x - \mu)/\sigma, \tag{2.1}$$

where μ is the mean and σ the standard deviation.

The problem of this method is that it uses estimators such as the mean and the standard deviation that have a very low breakdown point. Therefore, these estimators can be affected by the presence of only one extreme outlier. Thus, it can not only cause a masking problem but also influence the z-score of observations with normal behaviour (Seo, 2002).

Another limitation is that the z-score do not exceed 3 if the sample contains less than 11 observations, because the maximum z-score cannot exceed $(n-1)/n^{1/2}$ (Shiffler, 1988). For most of our dataset this method should not be effective.

2.4.1.2 Robust Z-score

Goktug *et al.,* (2013) performed an essay exposing a huge variety of normalization techniques that are a more robust alternative for the traditional z-score. The adaptation with more relevance in the literature was proposed by Iglwics & Hoaglin (1993). These authors proposed an adaptation of this method by using more robust estimators such as the median and median absolute deviation (MAD), that make this method more robust to the presence of outliers (Goktug *et al.,* 2013). The breakdown point of median is 0.5, which is quite high, so it would take more than 50% of outliers in the sample for these estimators to be affected. This adaptation of the z-score is effective with less observations than the traditional methods.

$$MAD = median\ (|x_i - median(x)|). \tag{2.2}$$

Generally, the rule used to classify an observation as possible outlier is when $|\ Z_i| > 3.5$ (Iglwics & Hoaglin, 1993). The MAD adaptation is defined as:

$$Z_{MAD} = (x - median)/MAD. \tag{2.3}$$

Nonetheless, in some practical applications it can be verified a very high presence of observations equal to zero, which is recurrent in accounting data due to the existence of accounts records without real applicability in some companies, or simply the existence of an attribute with median equals zero. To this purpose, which the MAD adaptation cannot be applied, the IBM (International Business Machines) [2] proposes a variation of the previous

method by replacing the MAD estimator for the mean absolute deviation (MeanAD), that is defined as [3]:

$$MeanAD = mean\ (|x_i - median(x)|). \tag{2.4}$$

And the IBM proposed adaptation as:

$$Z_{MeanAD} = (x - median)/MeanAD. \tag{2.5}$$

In Table 1 is represented the comparison of the different z-score adaptations and their influence in the score definition.

Table 1: Computation of different z-score adaptations.

i	*xi*	*yi*	*wi*	Z-Score (*xi*)	Z-Score (*yi*)	Z-Score (*wi*)	Z-MAD (*xi*)	Z-MAD (*yi*)	Z-MAD (*wi*)	Z-MeanAD (*xi*)	Z-MeanAD (*yi*)	Z-MeanAD (*wi*)
1	12	12	0	-0,290	-0,319	-0,289	-0,607	-0,569	-	-0,003	-0,004	0,000
2	6000	6000	0	3,175	3,160	-0,289	1815,530	1361,534	-	9,551	8,700	0,000
3	12	12	0	-0,290	-0,319	-0,289	-0,607	-0,569	-	-0,003	-0,004	0,000
4	13	13	0	-0,289	-0,319	-0,289	-0,303	-0,341	-	-0,002	-0,002	0,000
5	14	14	0	-0,289	-0,318	-0,289	0,000	-0,114	-	0,000	-0,001	0,000
6	15	15	1000	-0,288	-0,318	3,175	0,303	0,114	-	0,002	0,001	9,575
7	14	600	0	-0,289	0,022	-0,289	0,000	133,185	-	0,000	0,851	0,000
8	13	13	0	-0,289	-0,319	-0,289	-0,303	-0,341	-	-0,002	-0,002	0,000
9	12	12	0	-0,290	-0,319	-0,289	-0,607	-0,569	-	-0,003	-0,004	0,000
10	16	16	0	-0,287	-0,317	-0,289	0,607	0,341	-	0,003	0,002	0,000
11	17	17	0	-0,287	-0,316	-0,289	0,910	0,569	-	0,005	0,004	0,000
12	15	15	0	-0,288	-0,318	-0,289	0,303	0,114	-	0,002	0,001	0,000

Advantages of z-score

1. It can compare raw scores from data with different scales;
2. It can be adapted with more robust estimators of central tendency and dispersion like the Median and MAD;
3. The adapted z-score will not be directly dependent on the number of observations, and it can be applied to a small sample of data;
4. The Z-MAD adaptation will not suffer from the masking problem as shown is Table 1. The score value will not be influenced by the presence of outliers.

Disadvantages of z-score

1. Despite the possible implementation of small data samples, the method is more robust for large samples.
2. It is a statistical method with resource on central measure estimators, data with small variation can wrongly classify a high percentage of outliers.

2.4.1.3 Boxplot

Extreme values are obvious outlier candidates (Laurikkala *et al.*, 2000). An interesting approach in univariate extreme value analysis is to use the boxplot or "box-and-whisker" (Tukey, 1977). Tukey contribution to this method was the use of a robust five-number summary statistics like the maximum, minimum, "upper" (UQ) and "lower" (LQ) quartiles and the "median". The use of interquartile range (IQR) as a measure of variability and the median as a measure of central location makes this method robust to observations that deviate substantially from the rest of the data (Turkey, 1977). In Figure 4 is illustrated the graphic and statistic elements used in a construction of a boxplot.

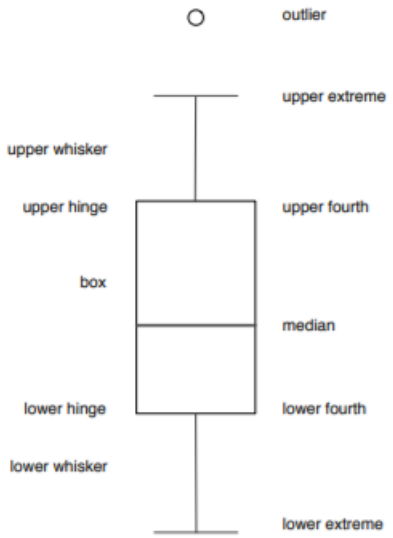

Figure 4 – Construction of a boxplot based on "five-number summary statistics" (Source: Wickham & Stryjewski, 2011).

The upper whisker is set at k times the IQR from the upper end of the box. Consequently, when there are no instances above k times the IQR of the upper quartile, the upper whisker is represented by the true maximum. The exactly equivalent rules hold true for the lower whisker (Aggarwal, 2017). John Tukey (1977) defines k as a non-negative value used to measures the range of the whiskers, where a value outside of k = 1.5 indicates a "moderate outlier" and k = 3 indicates a "far out". Therefore, the upper and lower whiskers of the boxplot can be defined by:

$$x_L = \max\{x_{(1)}, LQ - k * IQR\}, x_U = \min\{x_{(n)}, UQ + k * IQR\}. \quad (2.6)$$

2.4.1.4 Boxplot Adaptations

Wickham & Stryjewski (2011) have provided an extensive survey with the main extensions and variations since the "schematic plot" was introduced. There is a wide variety of graphic adaptations for this method, that provide more information in terms of the data density display such as the notched boxplot, vase plot, bean plot and violin plot. However, there are also variations for some basic definition such as the use different whiskers multipliers, fixed extreme quantiles or adjustments to the extremes considering the skewness.

In order to estimate the width of the central portion of the data distribution, the IQR can hardly be improved since it is the normal representation of half of the data distribution and its breakdown point is 25%. However, in tail area definition and anomalies in the data, a robust estimation of the boxplot extremes frequently needs an initial or auxiliary robust scale estimator (Shevlyakov *et al.,* 2013).

The *MAD* adaptation is frequently used because it has a simple and explicit implementation, needs a small computation time and is very robust. However, it does not have a high efficiency for normal data and it makes an implicit assumption about data symmetry, since it calculates the distance from a measure of central location (Rousseeuw and Croux, 1993; Shevlyakov *et al.,* 2013*).* Shevlyakov *et al.* (2013) propose other variations for the boxplot using other robust estimators like Q_n and S_n. These two scale estimators were proposed by Rousseeuw & Croux (1993). They are significantly more efficient alternatives to *MAD*. Unlike *MAD*, both statistics measure distances between values and not from a central location, so they do not depend on data symmetry. Both Q_n and S_n can be computed using O(n log n) time and O(n) storage. Their computer complexity is higher when comparing with *MAD* but have a better efficiency.

Interdecile Boxplot

In some practical applications a low data variability can be verified, thus 50% of the central data variability cannot be enough to define properly the whiskers range. When considering this dissertation practical application, an interesting boxplot adaptation is the use of the interdecile range instead of the IQR to find the width of the whiskers (Crone *et al.*, 2012). Similarly to the original method, it recurs to a set of non-parametric statistics called "Bowley's seven-number summary", which is an extension of the "five-number summary" with the addition of the *10th* and *90th* percentiles (Bowley, 1920). Therefore, the interdecile range is the difference between the *90th* and *10th* percentiles, so it is measure of 80% of data central

variability, which is more efficient for data with low variability. The rule that define the upper and lower whiskers for the interdecile boxplot is given by:

$$x_L = \max\{x_{(1)}, LQ - k * Int.R\}, x_U = \min\{x_{(n)}, UQ + k * Int.R\}, \quad (2.7)$$

where $Int.R$ is the interdecile range.

Advantages of Boxplot

1. It is a non-parametric method which makes no assumptions about the data distribution;
2. It is not dependent on low breakdown estimators, which makes it a robust method;
3. Has a fast and easy implementation, and the outputs are usually of easy comprehension;
4. The graphical representation shows the dispersion and skewness of the data, which can help when comparing multiple data sets;
5. Can be implemented for a large or small set of observations.

Disadvantages of Boxplot

1. Like the Z-score, it is a statistical method with resource on central measure estimators. Data with small variation can wrongly flag a high percentage of outliers;
2. Does not perform well on multimodal data or mixture distribution data;
3. The graphical representation does not display all the data.

2.4.2 Multivariate Machine Learning Algorithms

When considering the treatment of data in a multivariate space, an observation in a p-dimensional space, defined by multiple attributes, which is far from the rest of the data is considered a potential outlier. An observation may not be an outlier when his attributes are studied in a univariate context and still be an outlier in a multivariate context. This is explained due to the existence of observations with an unusual combination of attributes scores or by non-conformation with the correlation structure of the rest of the data (Jolliffe, 2002).

As stated previously in section 2.2.3, outlier detection is largely performed in unsupervised task, and due to the structure of our dataset we will only focus on this type of approach. In the literature, proximity-based methods have been addressed with different taxonomic composition. The categories and terms of the addressed methods vary for the different

authors. However, the properties of their taxonomy are basically the same. Therefore, this dissertation will follow the taxonomy proposed by Aggarwal (2017), which divides this category into: Cluster-based, Distance-based and Density-based algorithms.

Accordingly, all these classes of methods are related since they are based on a definition of proximity or similarity, which reproduces the strength of the relationships between two data instances (Aggarwal, 2017). In order to measure the similarity, it is required a measure distance. There is a wide variety of distance functions available in the literature, although their choice can be influenced by the type of data used. When dealing with continuous attributes the most popular choice is the Euclidean distance (Tan *et al.*, 2005). The Euclidean distance between the points *x* and *y* is defined as follow:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_j - y_j)^2}. \quad (2.8)$$

Proximity-based methods are very popular methods due to their intuitive simplicity and interpretability. They can be easily generalized to almost all types of data and were projected to detect both noise and outliers (Aggarwal, 2017).

Clustering-Based Methods

Clustering-based methods are one the most common techniques used in outlier detection, especially in the Financial fraud detection domain (Sabau, 2012; Ahmed *et al.*, 2016). The objective of these methods is to segment the data into meaningful homogeneous groups, being the members of different group dissimilar (Sabau, 2012). The three key assumptions of these methods are that outlier may not belong to any cluster, or belong to small or sparse clusters or lie far from the cluster *centroid* (Chandola *et al.*, 2009). There are many overlapping taxonomies of clustering algorithms, the most common and simple divide this class of methods in hierarchical-based, Partitional-based, Density-based, grid/graph based and model-based (Ahmed *et al.*, 2016). However, we will only discuss the most used in the literature.

Hierarchical clustering

These methods build a hierarchy of cluster based on two distinct strategies: agglomerative, when starting from singleton clusters, or divisive, when starting from a single cluster containing all observations. The most representative algorithms in the literature are BIRCH

(Balanced Iterative Reducing and Clustering using Hierarchies) (Zhang *et al.*, 1996) and CURE (Clustering Using REpresentatives) (Guha *et al.*, 1998). Both BIRCH and CURE are agglomerative algorithms that can recognize clusters with arbitrary shapes and are not sensitive to noise presence. However, BIRCH can deal with both categorical and numerical data and have a lower computer complexity.

Partitional clustering

These methods divide the data into a predefined set of partitions. They use an iterative procedure to move objects from a partition to another, towards improving the partition. Commonly, these methods operate with a centre-based cluster criterion. The centre of the cluster is a centroid, the average of all points within the cluster or a medoid, the most central point in the cluster (Ahmed *et al.*, 2016). The most popular partitional algorithms are K-means, PAM (Partitioning Around Medoids) (Kaufman & Rousseeuw, 1990), CLARA (Clustering LARge Applications) (Kaufman & Rousseeuw, 1990) and CLARANS (Clustering Large Applications based upon RANdomized Search) (Ng & Han, 2002).

Density-based clustering

This class of methods is probably the most popular in outlier detection. Moreover, considering they share properties from densities-based methods, they can deal with clusters with arbitrary shapes, sizes and densities. The most popular density-based cluster algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester *el al.*, 1996). The DBSCAN is the standard for this class of methods and adaptation of this method were proposed to deal with DBSCAN problem of find clusters with different densities. The most relevant are HDBSCAN (Hierarchical DBSCAN) (Campello *et al.*, 2015), OPTICS (Ordering points to identify the clustering structure) (Ankerst *et al.*, 1999) and SNN (Shared Nearest Neighbors) (Ertoz *et al.*, 2003). The main advantage of the presented methods is that they do not force all the instances to fit into a cluster, nonetheless their emphasis is still to find clusters.

Distance-Based Methods

Distance-based algorithms were presented by Knorr & Ng (1998). These authors considered the following notion of outlier: "An object *O* in a dataset *T* is a DB (*p*, *D*) outlier if at least fraction *p* of the objects in *T* lies greater than distance *D* from *O*". Ramaswamy *et al.* (2000),

extend this definition based on the full dimensionality distance between a point and their *k*th nearest neighbours and also provide a measure of outlierness to rank the outliers. Therefore, the assumption of this class of methods is that the *k*-nearest neighbours distance of an outlier is much higher (Chandola *et al.*, 2009). Distance-based methods perform a very detailed analysis with a significant computational cost, since they consider the entire granularity of a dataset (Aggarwal, 2017). Accordingly, the use of these methods is not recommended for very large datasets and they can struggle when the data contains both sparse and dense regions (Ramaswamy *et al.*, 2000; Breunig *et al.*, 2000). Another disadvantage of these methods is that they are only suitable for the identification of global outliers (Zhang *et al.*, 2007).

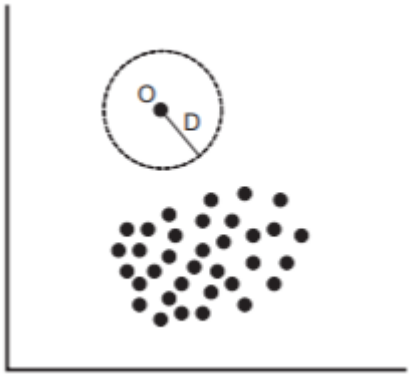

Figure 5 - An example of a distance-based outlier (Source: Zhang et al., 2007).

The most popular approaches in this class of methods are the *k*-nearest neighbours, hence they are non-parametric methods with an easy straightforward implementation. In these methods a wide variety of distance measures can be used to compute the *k*th nearest neighbours distance, nonetheless the distance can be measured using the k^{th}-nearest-neighbour (to single one) (Ramaswamy *et al.*, 2000) or the average distance for a number of *k*-nearest neighbours (Angiulli & Pizzuti, 2002). It is still worth mention the existence of parametric distance-based methods. The most popular are the Mahalanobis Distance, MCD (Minimum Covariance Determinant) and MVE (Minimum Volume Ellipsoid) (Finch, 2012).

Density-Based Methods

Density-based methods were introduced by Breunig *et al.* (1999,2000), which formally introduced the definition of local density and the Local Outlier Factor (LOF) algorithm. The key assumption in this class of methods is that "Outliers are points that lie in the lower local

density with respect to the density of its local neighbourhood" (Breunig *et al.*, 2000). Density-based methods solve the problem of Distance-based methods in the detection of local outliers, although they are still dependent on the computation of the full dimensionality distance between a point and its *k*-nearest neighbours (Zhang *et al.*, 2007). This class of methods is closely related to the clustering-based methods. Due to their notion of distance these methods are inter-dependent, while density-based methods partition the data-space, clustering-based methods partition the data points (Aggarwal, 2017). The LOF algorithm was developed based on some concepts such as "core distance" and "reachability distance" used for local density estimation, introduced in the early proposed density-based clustering techniques like DBSCAN and OPTICS (Breunig *et al.*, 1999). Many extensions and adaptation were proposed in the literature in order to improve and adapt LOF to new outlier detection context, the most notable extensions of this method are the LoOP(Local Outlier Probability) (Kriegel *et al.*, 2009), LOCI (Local Correlation Integral) (Papadimitriou *et al.*, 2003), COF (Connectivity-Based Outlier Factor) (Tang *et al.*, 2002), INFLO (Influenced Outlierness) (Jin *et al.*, 2006) and CBLOF (Cluster-Based LOF) (He *et al.*, 2003).

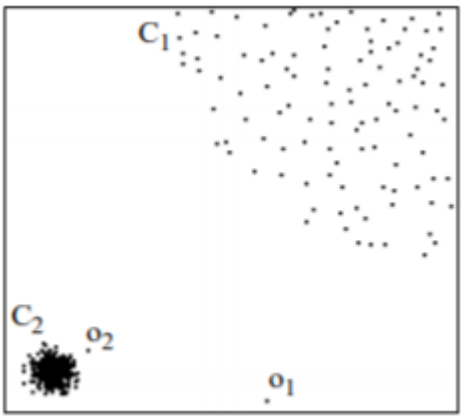

Figure 6 - Example of the impact of local density in a data set with different densities (Source: Breunig *et al.*, 2000).

2.4.2.1 LOF

Breuning *et al.* (2000), introduced the Local Outlier Factor (LOF), which is the standard density-based technique. The LOF algorithm indicates the degree of outlierness of each observation, based on the local density of an observation.

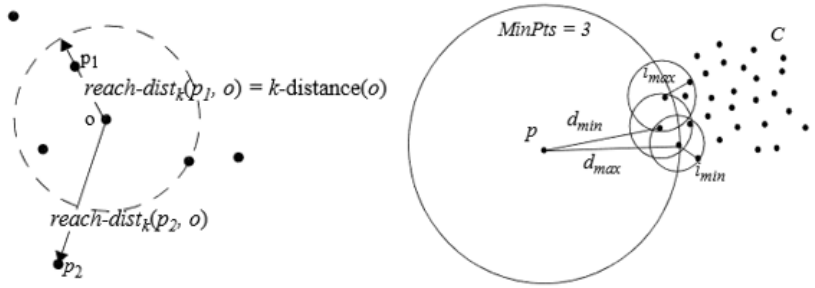

Figure 7— Illustration of reachability distance of point O and K-distance for a k=4 (left), and basic idea of LOF by comparing local density of a point with its neighbours, for *MinPts* =3 (right) (Source: Breuning *et al*., 2000).

Assuming that the *k*-distance of an object *p* is his distance to its *k*th nearest neighbours, all objects whose distance to *p* are not greater than *k*-distance set up the neighbourhood of object *p* (N_K). The reachability of two objects is the true distance between two objects. However, if the objects are "sufficiently" close, the true distance is replaced by *k*-distance of *O*. So, the reachability distance of an object *o* can be defined as:

$$reach-dist_k(p,o) = \max\{k-dist(o), d(p,o)\}. \tag{2.9}$$

To define the notion of density, LOF used the *MinPts* parameter for specify the minimum number of objects as a measure of volume, and use it to compare densities of a different set of objects in a dynamic way, determining the density in the neighbourhood of and object *p*. The local reachability density of an object *p is* defined as:

$$lrd_{MinPts}(p) = 1/\left(\frac{\sum_{o\in N_{MinPts}(p)} reach-dist_{MinPts}(p,o)}{|N_{MinPts}(p)|}\right). \tag{2.10}$$

Note that if the summation of all the reachability distances are 0, the local density can become infinite. This may occur if there are at least two different objects sharing the same spatial coordinates.

So, the Local Outlier Factor of an object *p* is defined as:

$$LOF_{MinPts}(p) = \frac{\sum_{o\in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}. \tag{2.11}$$

LOF is the average of the local reachability density ratio of object *p* from whose *MinPts* are nearest neighbours. If the LOF ratio is close to 1, indicates that the object is comparable to its neighbours, values below 1 indicates a dense region and significantly larger than 1 indicates an outlier.

Advantages of LOF

1. Designed to detect meaningful local outliers;
2. Easy implementation with only one parameter (*MinPts*);
3. Suitable for data with several clusters;
4. Easy to be generalized for different problems;
5. Output in form of observation score with intuitive interpretation.

Disadvantages of LOF

1. The LOF value varies non-monotonically for different *MinPts*;
2. There is no general definition for the value of LOF outlier score. The choice may consider the *MinPts* parameterization and the dataset density;
3. May struggle for data sets with different density regions;
4. High computational cost, because it must find all *MinPts* neighbours;
5. Do not specify why an outlier might be interesting.

2.4.2.2 DBSCAN

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was proposed by Ester *el al.* (1996), which is the standard density-based clustering technique. The key assumption of this method is that for a given *Eps* radius, each instance must contain a minimum number of *k* neighbours to form or belong to a dense region.

In this method the data instances are classified as core points, border points and outliers, using some definitions such as *Eps*-Neighbourhood, directly density-reachable, density-connected.

Eps-Neighbourhood – *Eps* (ε) represents the maximum radius distance threshold of a point *p* to its neighbours. *Eps*-Neighbourhood represent the instances inside of *Eps* radius from point *p*. A core point in a cluster is an instance that has at least a minimum number of instances (*MinPts*) in their *Eps* radius. The *Eps*-Neighbourhood can by defined as follows:

$$N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}. \quad (2.12)$$

Directly density-reachable – the instances that reside inside on the *Eps* radius of a point *q* are directly reachable from that point. Consequently, border points do not contain enough points in their *Eps* radius to form a dense a region but are directly reachable from a point

that belong to a cluster. Thus, clusters are formed around core points, and if two core points are directly reachable to each other, their clusters are merged. The directly density-reachability can by defined as follows:

$$p \in N_{Eps}(q) \; and \; N_{Eps}(q) \geq MinPts. \tag{2.13}$$

Density-reachable – when a point *p* is reachable from point *q* through a path of core points, being p_1,..., p_n, $p_1 = q$ such p_{i+1} is directly reachable from p_i.

Density-connected – when considering a point *o*, two points are density connected if they are both reachable from *o*.

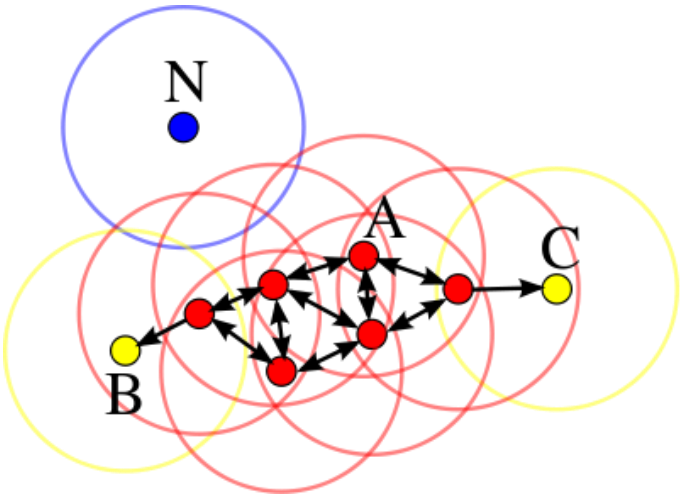

Figure 8 - Illustration of DBSCAN cluster model for *MinPts* =4 (Source: Schubert *et all.*, 2017).

In Figure 8, an example of DBSCAN implementation is shown for *MinPts*=4. The area surrounding the point represents the *Eps* radius containing a minimum of 4 points. All the points form a cluster because they are all density-connected and all density-reachable from any point in the cluster. The points B and C are border points because their *Eps*-Neighbourhood less than the *MinPts* but are directly reachable from a core point. The point N is a noise point since it is not directly reachable to any point in the cluster and its *Eps*-neighbourhood do not contain enough *MinPts* to form a new cluster region.

Parametrization

As stated above the DBSCAN implementation requires the specification of the *MinPts* and *Eps* parameters. The choice of the distance function, *dist(p,q)* can be considered a parameter because will also influence the shape of the neighbourhood.

The *MinPts* parameter should base on the problem application, the only restriction is that it must assume a value greater than one, to avoid that all points are classified as core points.

The value of *Eps* can be chosen based on *k-dist* graph, where $k = MinPts - 1$. Then by plotting sorted the *k-dist* values in ascending order, the graph of this function provides some insights about the data distribution in the data set. The threshold should be where the "valley" is formed in the graph.

The DBSCAN supports any distance function, which should be chosen appropriately by considering the data set properties. This parameter is impactful in the estimation of *Eps* radius.

Advantages of DBSCAN

1. Can find clusters with arbitrary shapes, even when surrounded by different clusters;
2. Only requires two input parameters and is mostly insensitive to the order of points in a database;
3. *MinPts* can be established by an expert application domain;
4. Has a concept of core and border points, and is robust to outliers;
5. Do not force all the points to be clustered;
6. The output is in discrete form with intuitive interpretation, where the outliers are assigned to cluster zero.

Disadvantages of DBSCAN

1. It is not effective for data with varying densities, because *Eps* will not have an appropriate value for the different regions;
2. It is not recommended for high-dimensional data. The algorithm computational complexity can ascend to $O(n^2)$;
3. The *Eps* is commonly established by a graph interpretation and this parameter is heavily influenced by the chosen distance function.
4. Do not specify why an outlier might be interesting.

Chapter 3 - Methodology

3.1 Data Description

The data set used in the practical implementation of this dissertation is composed of data collected through the Quarterly Survey on Non-Financial Corporation (ITENF). The ITENF is a statistical operation for collection of accounting attributes carried out by the BDP in cooperation with the Statistics Portugal (INE). This survey comes from the need of both organizations responsibilities to publish statistics of Non-Financial Corporation sector in a quarterly basis. In this statistical operation, approximately 4 thousand companies are surveyed. The data from ITENF is used in an extrapolation process that aims to infer the results for the total number of Portuguese Non-Financial corporations in terms of total assets and turnover. For that purpose, each company is assigned with an extrapolation factor that amplifies individual values of each company many times as necessary to reach the universe of companies in a given sector of activity. Another objective is to reduce the statistical burden on respondents. The ITENF data collection is regulated by "Decreto-Lei n.º 136/2012, de 2 de julho" and " Lei Orgânica do Banco de Portugal".

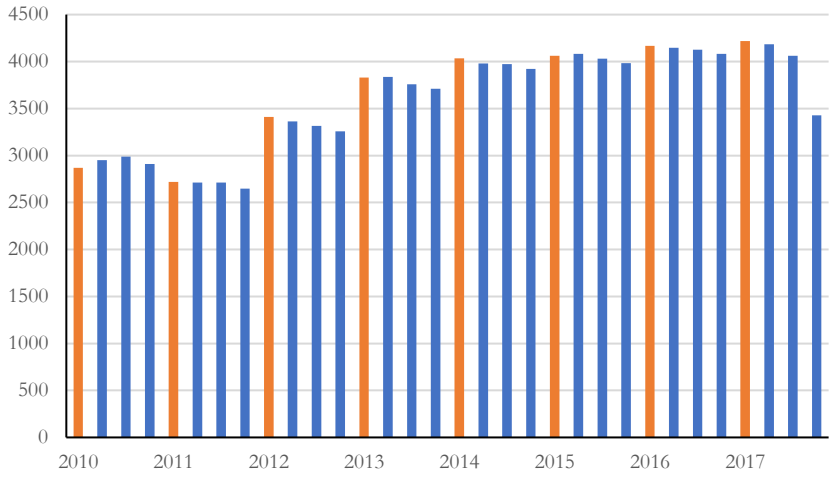

Figure 9- Number of companies answering ITENF.

The companies selected for the ITENF vary in each year, which means that there are companies with different amounts of observations. The differences between the sample size and the number of ITENF responses shown in Figure 9 is due to non-response by companies, mergers and divisions. However, the group of companies with high dimension are always included in the surveys due to their representativity in the business universe.

The data set have a total of 115.473 observations collected between the period from 2010 to 2017. Although BDP and INE joint survey was initiated in 2000, until 2009 the survey followed the accounting principles of the Official Accounting Plan (POC), regulated by "Decreto-Lei n.º 410/89, de 21 de novembro". From 2010 onwards, the POC was replaced by the Generally Accepted Accounting Principles (SNC), through "Decreto-lei n.º 158/2009 de 13 de julho. The change in accounting regulations lead to the creation, removal and aggregation of certain accounting attributes, making the comparison of companies in both regulations unreasonable.

In Figure 10 it is shown an extract of the dataset. It is verified that the companies have a panel data structure, where each entity has multivariate observations for quarterly periodicity.

CompanyID	PeriodID	extr.fact	CAE	Size	B05	B10	B15	B25	B30	B41	B51
1120000034	31/03/2015	4,625836	42	3	231000	1854091	0	1311614	0	969514	605020
1120000034	30/06/2015	4,640663	42	3	428000	1822110	0	1690262	0	761827	555353
1120000034	30/09/2015	4,650457	42	3	907200	2548774	0	2052287	0	375927	401647
1120000034	31/12/2015	4,592556	42	3	1040778	3140316	0	2582866	0	414312	609013
B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	
1410959	1410959	0	0	0	53344	2062894	0	271822	0	0	
1382235	1239299	0	0	0	53344	2062822	128760	276899	0	0	
1097335	1097335	0	0	0	53344	1312861	128760	279299	0	0	
954864	954864	0	0	0	53266	1244751	126829	232897	0	0	

Figure 10 – Extract from the data set relative to company ID nº- 1120000034.

The data set contains 10.028 companies, which are all companies so far selected for ITENF. In this survey 93 attributes are collected, most of them related to the balance sheet (BS) and income statement (IS). The IS attributes are cumulative and have an inconsistent behaviour, since they represent the everyday company activity. In other hand, BS attributes are more constant, since they represent the company's patrimonial position over time. Due to the limited time and resources, the practical implementation will only focus on 18 attributes from BS, selected by the analysts due to their relevance in the production of statistical indicators.

Table 2: Description of the characterization and Balance sheet attributes.

Characterization			
Attribute	Type	Cod Var.	Description
CompanyID	Integer	-	Code assigned to each company in order to anonymize the database.
PeriodID	Date	-	Quarterly data.
Extrapolation factor	Continuous	-	Number of companies in the population that are represented by each of the companies observed in the sample.
CAE	Discrete	-	Portuguese Classification of Economic Activities (CAE) regulated by "Decreto-Lei nº 182/93 de 14 de Maio".
Size	Discrete	-	Dimension of the company, which is defined by the four categories (micro, small, medium and large entities) described in SNC.
Balance Sheet			
Attribute	Type	Cod Var.	Description
Other financial instruments	Continuous	B05	Intended to include financial instruments that are not recognized in the other attributes.
Trade debtors:	Continuous	B10	Amounts owed by the customers for purchases of goods and services made on credit.
-Non-residents trade debtors	Continuous	B15	Amounts owed by non-resident customers for purchases of goods and services made on credit.
Trade creditors:	Continuous	B25	Amounts owed to suppliers for purchases made on credit.
-Non-residents trade creditors	Continuous	B30	Amounts owed to non-resident suppliers for purchases made on credit.
Other accounts receivable	Continuous	B41	Assets that are not recognized in the other attributes.
Other accounts Payable	Continuous	B51	Liabilities that are not recognized in the other attributes.
Obtained funding:	Continuous	B60	Interest-bearing liabilities
-Borrowings from Financial institutions	Continuous	B65	Borrowing from a financial institution.
-Bonds issued	Continuous	B70	Debt issued through bound.
-Equity investors	Continuous	B76	Interest-bearing borrowings from shareholders.
-subsidiaries, assoc. and joint ventures	Continuous	B78	Borrowings from subsidiaries, associates and joint ventures.
Shareholders	Continuous	B82	Net receivables from shareholders.
Financial Investments	Continuous	C50	Long-term equity investiments.
Investment Properties	Continuous	C60	Lands or building held to earn rentals or for capital apreciation.
Tangible fixed assets	Continuous	C75	Lands and buildings, vehicles, machenary and other tangible goods whose function is to produce or supply goods and services.
Intangible assets	Continuous	C80	Assets whithout physical substance, such as trademarks, patents and software.
Non-current assets held for sale	Continuous	C90	Non-current assets (or disposal groups) held for sale rather than for continuing use in production.

Table 3: Frequency table for the number of observations and companies in the data set.

Nb Obs:	Nb Companies:	Total Obs.	Freq.Rel (companies)	Freq.Rel (obs)	Freq.Rel.Ac (companies)	Freq.Rel.Ac (Obs)
1	267	267	2,66%	0,23%	2,66%	0,23%
2	286	572	2,85%	0,50%	5,51%	0,73%
3	434	1302	4,33%	1,13%	9,84%	1,85%
4	3277	13108	32,68%	11,35%	42,52%	13,21%
5	69	345	0,69%	0,30%	43,21%	13,50%
6	87	522	0,87%	0,45%	44,08%	13,96%
7	227	1589	2,26%	1,38%	46,34%	15,33%
8	1228	9824	12,25%	8,51%	58,59%	23,84%
9	46	414	0,46%	0,36%	59,04%	24,20%
10	62	620	0,62%	0,54%	59,66%	24,74%
11	126	1386	1,26%	1,20%	60,92%	25,94%
12	684	8208	6,82%	7,11%	67,74%	33,04%
13	20	260	0,20%	0,23%	67,94%	33,27%
14	35	490	0,35%	0,42%	68,29%	33,69%
15	123	1845	1,23%	1,60%	69,52%	35,29%
16	540	8640	5,38%	7,48%	74,90%	42,77%
17	20	340	0,20%	0,29%	75,10%	43,07%
18	35	630	0,35%	0,55%	75,45%	43,61%
19	89	1691	0,89%	1,46%	76,34%	45,08%
20	437	8740	4,36%	7,57%	80,69%	52,65%
21	21	441	0,21%	0,38%	80,90%	53,03%
22	44	968	0,44%	0,84%	81,34%	53,87%
23	101	2323	1,01%	2,01%	82,35%	55,88%
24	425	10200	4,24%	8,83%	86,59%	64,71%
25	19	475	0,19%	0,41%	86,78%	65,12%
26	30	780	0,30%	0,68%	87,08%	65,80%
27	71	1917	0,71%	1,66%	87,78%	67,46%
28	320	8960	3,19%	7,76%	90,98%	75,22%
29	38	1102	0,38%	0,95%	91,35%	76,17%
30	42	1260	0,42%	1,09%	91,77%	77,26%
31	146	4526	1,46%	3,92%	93,23%	81,18%
32	679	21728	6,77%	18,82%	100,00%	100,00%
Total	**10028**	**115473**	**100,00%**	**100,00%**	**100,00%**	**100,00%**

Table 4: Descriptive statistics of Balance Sheet attributes.

Cod Var.	Min	1º Quart.	Median	Mean	3º Quart.	Max	Std
B05	-251.029.000	0	0	434.894	0	1.091.400.313	10.684.958
B10	-274.368.761	167.830	1.266.425	7.312.874	6.049.107	735.612.023	25.004.463
B15	-151.840.803	0	448	2.146.354	657.497	481.918.269	11.778.732
B25	-26773048	113.588	797.557	6.118.133	4.013.740	1.060.756.993	27.087.228
B30	0	0	9.245	1.787.210	527.816	750.000.000	11.546.992
B41	0	7.724	236.683	7.572.644	1.764.027	2.859.680.169	63.323.185
B51	0	51.741	492.664	7.840.601	300.576	2.149.996.711	55.132.048
B60	0	18.750	867.292	32.842.058	7.986.302	15.134.119.158	280.912.295
B65	0	0	214.260	11.149.222	3.354.742	1.306.028.846	55.132.048
B70	0	0	0	8.931.788	0	13.009.142.000	168.252.028
B76	0	0	0	8.529.559	0	6.339.790.616	114.426.423
B78	0	0	0	2.571.585	0	441.770.182	64.993.417
B82	-613.633.899	0	0	5.389.272	0	8.844.025.000	133.423.098
C50	0	0	2.279	25.498.053	427.816	16.219.639.555	295.012.834
C60	0	0	0	1.560.167	0	768.678.198	18.543.031
C75	0	72.838	781.040	12.147.771	4.820.856	4.874.493.304	89.400.170
C80	0	0	193	9.639.071	49.059	28.768.164.970	200.915.568
C90	0	0	0	119.515	0	587.259.327	4.886.293

Through the analysis of the Tables 3 and 4 it is possible to identity that the biggest challenge in this dissertation will be due to the composition of the dataset. Thus, some of most important aspects and limitations of the data set are:

1. About 46% of companies have less than 8 observations (2 years);
2. At the observations level, about 55% of the data set is represented by companies with at least 20 observations (5 years);
3. The most common is a company only be surveyed for a year, represent around 33% of the dataset;
4. Large companies are always surveyed based on the European commission criterion, they represent around 7% of the companies and 19% of the dataset;
5. The existence of companies with no periodicity of 4 observation (4 per year), is an indicative of missing period due to non-reporting or other special events, like mergers and splits;
6. Although the data have a panel structure, it is not possible to apply time series methods considering the small number of observations per company, it would not be possible to satisfactorily extract the trend and seasonal components from time series;
7. Based on the descriptive statistics Table 3, it is verified that a high number of attributes are reported as zero by the companies, which in accounting means that these attributes do not have expressivity in the activity of the company;

3.2 Data Pre-processing

Data cleaning

The data provided by the BDP has already been submitted to some processes of quality control. Nevertheless, does not mean that the data is prepared for the practical implementation. After an exploratory analysis of the data, it was possible to identify observations with missing and inaccurate reports. The missing report are observations where all the attributes were reported as zero, which can be a consequence of the absence of activity in the company, state of insolvency or non-reporting. The inaccurate reports are observations with extrapolation factor equals zero, which means that these observations are not used for statistical purposes. Notwithstanding, companies with less than 4 observations are not also in the interest of analysis, since it is not possible to perform any significant analysis with such a reduced number of observations. Therefore, it was removed from the dataset 3.952 observations relative to missing and inaccurate reports and 2.244 observations relative to

companies with less than 4 observations. A new frequency table for the number of observations per company in the dataset is presented in Table 17, Annex A.

Normalization

In this dissertation, only continuous attributes will be analysed. As shown in table 4, multiple attributes in a dataset are expressed in different scales. To avoid the predomination of certain attributes in a multivariate analysis, it is recommended the transformation of the data using normalization. The most common normalization methods are the min-max normalization, where the observations are normalized into a common interval [min,max], and standardization, where the observations are transformed making all attributes have mean equals 0 and standard deviation equal 1. Frequently, it is desirable to standardize over normalize when dealing with outliers, since the scale and dispersions measures are maintained, as well as the correlations between attributes. The standardization was already presented as Z-score method (section 2.4.1.1, equation 2.1) (Gama *et al.*, 2010).

3.3 Methods Implementation/Parametrization

It is not an easy task to select a proper subset of methods for this work keeping in mind that several algorithms have been proposed. Nonetheless, the chosen methods were based on the specification of the dataset and the availability of implementation in CRAN package repository.

The first objective in this dissertation was to find outliers in the univariate and multivariate space. In the univariate space it was used statistical methods, most of these methods are limited by the knowledge of the underlying distribution of the data and by the number of observations in the dataset. The choice of Z-score and Boxplot is justifiable by their applicability for small samples and are also non-parametric methods which makes no assumption about the underlying distribution. Another justification is their easy implementation and interpretability of results. The R work directory already support a function for the boxplot implementation based on Tukey (1976).

In the multivariate space it was decided to implement the most popular cluster and density-based methods in the literature, which are the LOF and DBSCAN. The choice of these methods is mainly justified by the lack of labels in the dataset and also because of its composition, given that it is composed by an aggregate of companies with distinctive behaviours, which must be represented by several density regions. Another important aspect

is that the algorithms do not require many input parameters and the output has intuitive interpretation, which is significative for the BDP future implementations. Due to the algorithm popularity, there are several packages in the CRAN repository with the implementation of this methods. The packages used in this dissertation are as follows:

"dprep" - It was created by Acuna *et al.* (2005) for data pre-processing and visualization. It supports function for normalization, treatment of missing values, discretization, feature selection and outlier detection. For outlier detection it supports the lofactor function. This function finds de local outlier factor for each observation based on Breuning *et al.* (2000) proposed method.

"dbscan" – It was created by Hahsler *et al.* (2017), has several fast implementations of density-based algorithms of DBSCAN family for spatial data, such as DBSCAN, OPTICS, HDSCAN and LOF. It also provides a SNN clustering implementation and a fast calculation of the *k*-nearest neighbours distances in a matrix of points.

3.3.1 Z-score

The first implemented method in this dissertation is the Z-score method. The choice of this method is exclusively related with the small number of observations required for his application (section 2.4.1.2). Thus, the Z-score will be calculated for each company attribute in the dataset. To optimize the results and minimize de number of false outliers detected, its implementation should not be straightforward. In the first step, it will be excluded from the analysis attributes with a range of values lower than 20.000 and with a coefficient of variation (CV) lower than 30%. The range of values parameter was set by the analysts, considering that attributes with low values will have no impact on statistical studies. The CV parameter will determine if there is not a great variability of the data relative to the location of the middle of the distribution. So, for a low CV the data tend to be more stable and will be improbable the presence of large swing in the data. Therefore, this parameter will determine if the attribute present a significant variation that justifies the Z-score implementation, since if that hold not true, observations with meaningless variations would be flag as outliers. The CV is defined as follow:

$$CV = \sigma/\mu. \tag{3.1}$$

The second step will be the calculation of *MAD*. As stated previously (section 2.4.1.2), different robust Z-scores will be implemented based on the *MAD* value, Z_{MeanAD} when

MAD equal zero and Z_{MAD} when different from zero. In the third and last step, will be flag as outlier observation with an absolute score higher than 3.5 (Iglwics and Hoaglin, 1993).

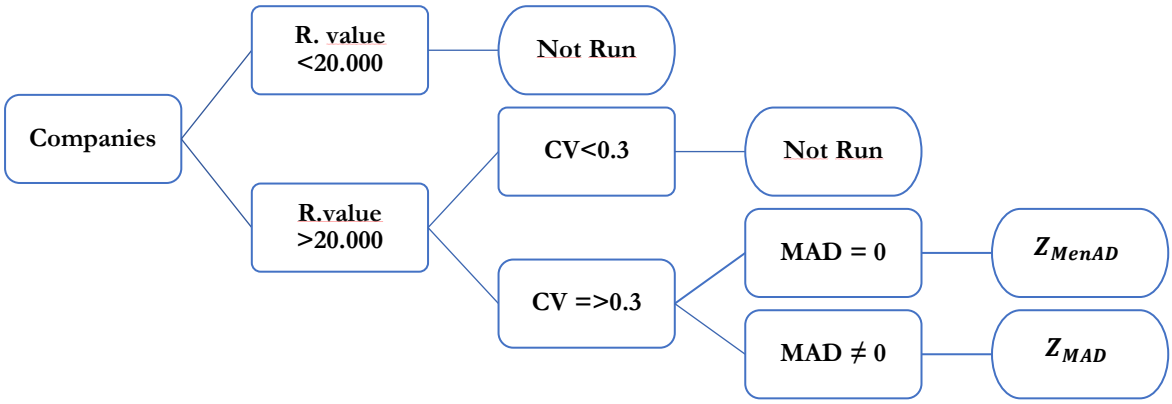

Figure 11 - Implementation structure of Z-score method.

3.3.2 Boxplot

The Boxplot is a univariate non-parametric statistical graphical representation that will be performed for each company attribute. Like Z-score, Boxplot will not have a straightforward implementation. The boxplot can be applied to a small number of observations. However, it was not found an author that defines the least number of observations for which a boxplot should be appropriate. Despite a boxplot could be constructed with a minimum of 5 observation ("five-number summary"), the larger the sample size the meaningful the quartile definition. In this way, it was defined that this graphical representation will only be implemented for companies with at least 20 observations. As for Z-score, will be excluded from the analysis attributes with a range of values lower than 20.000.

One of the problems detected during the first approach to this method was that for attributes with little or no central variability the IQR value would be too low for the whisker to be properly defined. Thus, to ensure that the central portion of the data has enough variability, an adapted coefficient of variation (CVA) was used. This uses more robust estimators such as the IQR as a measure of dispersion and de median as a measure of central location (Murphy *et al.*, 1998). The CVA can be defined as follow:

$$CVA = IQR/Median. \quad (3.2)$$

Like in Z-score for the boxplot to be implemented the CVA should be at least 30%. When this condition is not verified an interdecile boxplot is applied to solve the problem of low

central variability of the data, since it defines the whiskers length based on 80% of the data variability.

Nonetheless, when the median of an attribute is zero the CVA cannot be implemented. A zero median can be caused by a percentage of zeros higher than 50% or by zero being the value in the centre of the spectrum of values. When the percentage of zeros is higher than 50%, there are no central variability of the data and the whiskers cannot be defined, so an interdecile boxplot is applied. When zero in the centre of the spectrum of values a regular boxplot is applied.

The boxplot parameter K was set to 3, which was the value set by Tukey (1976) to define an observation as a "far out".

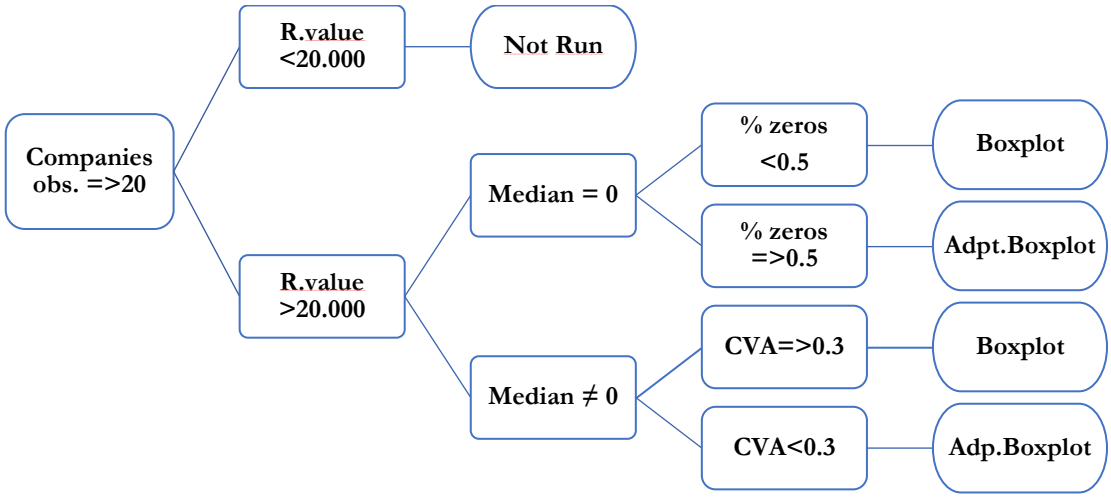

Figure 12 - Implementation structure of Boxplot method.

3.3.3 LOF

In the LOF implementation, the only a parameter that needs to be specified is the *MinPts*. It represents the number of *k*-neighbours for local density definition. The generated scores are very sensitive to the *MinPts*. Unfortunately, the LOF scores do not decrease or increase monotonically. In order to get stabilized LOF, it was proposed in the original paper an approach to define a lower and upper bounds for *MinPts* definition, and for each point keep the maximum LOF value of *MinPts* range.

The minimum *MinPts* should represent the least size for a group of N points to be considered a "cluster" or dense region. The minimum *MinPts* will be set to 3, which corresponds to the annual sample. The 3 closest observations should be the closest reported periods by the

company. Another explanation is the fact of 4 being the companies with the least number of observations considered in the practical implementation. In cases where *MinPts* is higher than companies sample size, the local density will have in consideration observations from different companies.

The maximum *MinPts*, should be the maximum number of objects to be considered outlier when clustered together. This will also allow to search outliers for the companies based on its usual behaviour. The maximum *MinPts* will be set to 19, so a dense region could be set by group of 20 observation (5 years). This is also justifiable by sample size chosen in Boxplot, being able to perform a direct comparation of results.

In addition, the interpretability of LOF score is also an important aspect in the analysis. Typically, values close to one indicate are inlier observations (not outlier), but there is no specific score to classify an outlier. The output score is sensitive to the parametrization and to the local density fluctuation in the dataset. Therefore, the outlier score was defined through a visual analysis to the increasingly ordered LOF scores, a cut was made in the plot "valley", which is where it is verified the greater variability of the values. We also had in consideration that the outliers are a rare phenomenon, so the percentage of outliers should not be high.

3.3.4 DBSCAN

In DBSCAN it is only required the definition of two parameters, *MinPts* and *Eps* radius. Although, the *MinPts* of DBSCAN is slightly different from LOF. In DBSCAN the *MinPts* is the number of points within *Eps* radius for a point to be considered a core point and define a new density zone. Therefore, for the same reason used in LOF the *MinPts* will be set to 4. The closest 4 observations, which correspond to the annual sample, it will form a dense region or belong into the same cluster. For companies with more than 20 observations, it will also be implemented a *MinPts* of 20. The objective is to analyse the results for a different level of granularity. Looking not only to observations that vary in relation to their closest reports but also in relation to the usual behaviour of the company. This will also provide a term of comparability between Boxplot and LOF for *MinPts* equal to 19.

The *Eps* parameter is influenced by the choice of *MinPts*. Thus, the *Eps* parameter will assume different values for *MinPts* equal to 4 and 20. As stated in section 2.4.2.2, the *Eps* can be

chosen based on *k-dist graph* where $k = MinPts - 1$, ordered in an ascending order. The threshold should be where the "valley" is formed in the graph.

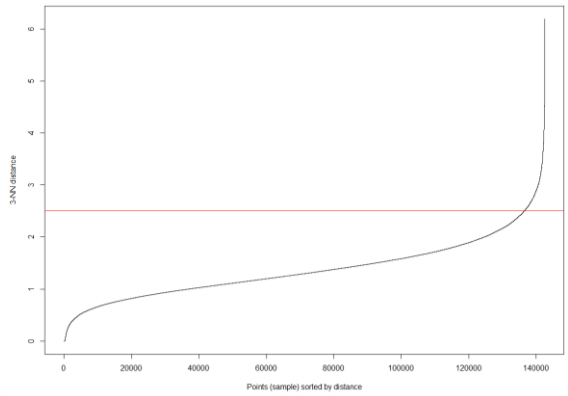

Figure 13 – K- distance graph for k equal 3, and with cut off at 2.5

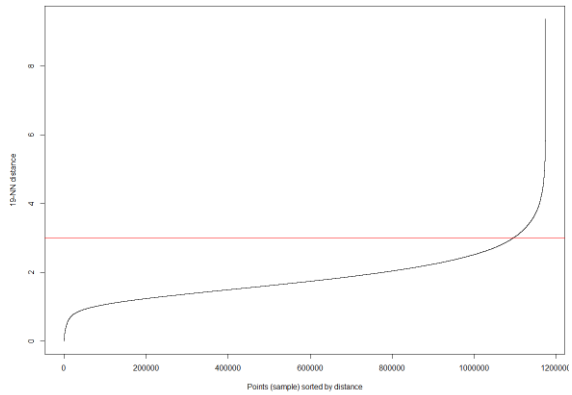

Figure 14 - K- distance graph for k equal 19, and with cut off at 3

As shown in the Figures 13 and 14, for *MinPts* equal 4 the *Eps* radius is 2.5 and for *MinPts* equal 20 the *Eps* radius is 3. The choice was made through visual analysis, which is influenced by the visual perspective of the analyst.

Chapter 4 - Results

In this chapter will be presented and discussed the results for the univariate and multivariate methods. This division will consider the level of granularity of the methods to facilitate the comparison between them. For the univariate methods it is possible to identify why the outliers might be interesting, since it is possible to identify the cause of the outliers, allowing to perform a more detailed analysis to the results. In other hand, multivariate methods identify outliers at the observation level, which requires a technical analysis to understand its causes.

4.1 Univariate Results

At the univariate level will be presented the results of the z-score and boxplot method. In the analysis will be considered the number of outliers detected by attributes, CAE, size of the companies and quarterly period.

The analysis performed at the attributes level is represented in the Tables 5, 6 and 7. These tables represent the number of outliers detected in an attribute of a certain company. As in table 5 is represented the number of outliers detected through the z-score method for companies with less than 20 observations for exactly the attribute B05 were identified 321 companies with only one outlier and 78 companies with two outliers. Thus, it will be possible to know which attributes are most and less affected the presence of outliers and to understand the behaviour that leads to these results.

Table 5: Number of outliers detected per attribute with z-score method for companies with less than 20 observations.

Nb Out	B05	B10	B15	B25	B30	B41	B51	B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	Total Var.	Total (%)	Total Obs.
1	321	756	728	848	809	1113	1091	568	513	39	385	87	706	246	124	338	214	43	8929	59,85	8929
2	78	194	207	226	214	344	287	167	166	26	94	27	191	73	39	118	55	10	2516	16,86	5032
3	61	110	107	115	119	224	171	124	110	12	88	11	157	96	34	134	99	10	1782	11,94	5346
4	16	48	56	53	51	95	114	63	58	3	46	4	67	62	8	59	52	5	860	5,76	3440
5	16	19	19	27	27	52	45	34	31	1	23	1	36	58	14	42	36	3	484	3,24	2420
6	4	7	12	6	14	14	14	12	10	1	3		12	17	3	12	8	1	150	1,01	900
7	8	4	5	5	5	15	6	16	12	2	12	1	9	19	4	17	16		156	1,05	1092
8	2		2		1	2	2	2	1		3		4	4	1	4	2		30	0,20	240
9		1					1	2	2		1		2			2	1		12	0,08	108
Total	506	1139	1136	1280	1240	1859	1731	988	903	84	655	131	1184	575	227	726	483	72	14919	100	27507

Table 6: Number of outliers detected per attribute with z-score method for companies with at least 20 observations.

Nb Out	B05	B10	B15	B25	B30	B41	B51	B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	Total Var.	Total (%)	Total Obs.
1	164	227	215	256	292	313	286	143	187	34	132	79	257	82	45	73	115	59	2959	23,11	2959
2	123	142	166	199	210	210	159	123	135	29	107	54	205	53	18	58	80	20	2091	16,33	4182
3	95	96	142	126	142	157	153	93	110	38	97	64	200	67	54	94	115	24	1867	14,58	5601
4	75	83	84	77	97	215	278	87	96	42	122	30	158	63	22	51	86	22	1688	13,18	6752
5	31	38	54	44	69	122	121	59	68	18	57	25	89	58	16	59	72	10	1010	7,89	5050
6	27	27	34	43	34	85	82	43	43	11	39	12	50	43	17	31	46	3	670	5,23	4020
7	18	22	31	35	46	58	61	39	46	8	18	8	55	45	16	36	62	3	607	4,74	4249
8	18	17	28	19	20	55	28	32	30	8	18	4	27	43	6	22	37	3	415	3,24	3320
9	9	16	26	12	23	34	25	32	34	2	17	5	39	40	8	29	30	1	382	2,98	3438
10	17	15	11	11	11	18	14	24	25	7	12	1	30	36	1	15	28	2	278	2,17	2780
11	14	8	12	5	14	20	11	26	21	8	16	3	21	35	10	11	38	2	275	2,15	3025
12	15	3	7	1	10	4	8	14	21	2	10	2	16	21	3	10	15	1	163	1,27	1956
13	10	7	6	2	7	3		20	15	5	13	2	15	33	7	12	17	6	180	1,41	2340
14	6	2	6	1	3	3	1	10	17	6	9		14	26	4	5	6		119	0,93	1666
15	4		8		4	5		14	8	1	5	4	11	15	5	5	9	1	99	0,77	1485
Total	626	703	830	831	982	1302	1227	759	856	219	672	293	1187	660	232	511	756	157	12803	100	52823

Table 7: Number of outliers detected per attribute with boxplot for companies with at least 20 observations.

Nb Out	B05	B10	B15	B25	B30	B41	B51	B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	Total Var.	Total (%)	Total Obs.
1	149	233	220	263	258	259	246	171	178	34	132	71	224	115	63	140	132	56	2944	40,26	2944
2	109	101	115	84	133	149	127	113	123	32	95	56	204	95	28	78	91	28	1761	24,08	3522
3	106	39	91	40	89	85	53	101	141	33	114	67	194	103	71	72	123	29	1551	21,21	4653
4	60	15	35	22	35	45	18	46	46	12	52	25	97	47	13	16	49	16	649	8,87	2596
5	18	2	13	11	15	16	14	15	18	2	3	3	35	16	3	14	20	1	219	2,99	1095
6	12	5	4	5	4	9	3	6	10	1	1	1	24	14		2	13	1	115	1,57	690
7	7	1	4	3	4	2	3	2	5				9	3			9	1	53	0,72	371
8	2	3		1									4						10	0,14	80
9	3			1									5						9	0,12	81
10	1																		1	0,01	10
11		1																	1	0,01	11
Total	467	400	482	430	538	565	464	454	521	114	397	223	796	393	178	322	437	132	7313	100	16053

In the Table 5 is presented the number of outliers detected per attribute with z-score method for companies with less than 20 observations. It was detected a total of 27.507 outliers, representing around 3.21% of the total univariate observations. There were identified outliers in 14.919 attributes, which represent 13.54% of the total attributes. As expected, the majority of the detected outliers are extreme isolated single values, since around 60% of the identified cases only one outlier was flagged. The maximum of 3 outliers in an attribute represent around 89% of the detected cases.

In the Table 6 is presented the number of outliers detected per attribute with z-score method for companies with at least 20 observations. It was detected a total of 52.823 outliers, representing around 4.75% of the total univariate observations. There were identified outliers in 12.803 attributes, which represent 30.76% of the total attributes. The detection of extreme isolated single values also prevailed in the results, but only represent 23% of the cases. Attributes with few outliers detected have less significance, since for a maximum of 4 outliers in an attribute only 67% of the cases are represented. For attributes with more observations, it was expected that the data also have more volatility, because they reflect the evolution of the corporate behaviour over a longer period, meaning that more outliers per attribute tend to be identified due to their higher variability.

In the Table 7 is presented the number of outliers detected per attribute through the boxplot for companies with at least 20 observations. It was detected a total of 16.053 outliers, representing around 1.44% of the total univariate observations. There were identified outliers in 7.313 attributes, which represent 17.57% of the total attributes. Like z-score, the extreme isolated single values prevailed in the results and represent 40% of the cases. Nonetheless, attributes with few detected outliers, a maximum of 4 in an attribute, represents around 95% of the cases. Boxplot is a more robust method than Z-score. So it was expected a smaller number of outliers being identified, since this method is non-parametric and tends to deal better with high variability data, flagging less possible false outliers.

However, there are cases where a large number of outliers were detected per attribute. These reflect the cases where the implementation of the methods is not suitable. In z-score are usually associated with attributes with high variability, non-normal distribution, multimodal data and temporary shifts (Annex B). In boxplot are usually associated with arbitrary behaviour and temporal shifts (Annex C). Both methods do not take into consideration the temporal component of the data, so when a temporary or permanent shift of the attribute

values is verified multiple outlier tend to be flag, which in a contextual time series method only count as one. In the Annex D are represented Tables 18, 19 and 20 with the higher sequence of outliers detected for the number of outliers detected in the attributes. Generally, a high number of outliers verified in an attribute is associated with a large sequence of outliers detected.

When analysing the total number of outliers detected for the different attributes, it is verified some similar results in the different tables. As shown, in all tables the attributes with less outlier detected were the attributes B70, B78, C60 and C90. These attributes tend to be stable or non-existent in several companies, especially in small firms. The B70 and B78 are both related with the companies obtained funding. The most common form of funding is through financial institutions. So, funding operations though bond market (B70), such as commercial paper or company-issued bonds, are generally practiced by large companies. The B78 is associated to funding from entities in which the company as some influence/control, such as associates. So, only companies with shares in other companies will have movements in this attribute. The C60 are investment properties usually related to large or real estate companies, since this attribute represent properties held for rental or capital appreciation. The C90 represent shared companies held for sale or disposal.

The attributes with more outliers detected in both z-scores tables were the B41, B51 and B82. The B41 and B51 are attributes directly associated with several IS attributes. These are also directly affected by the volatility and seasonality of the company´s activity. These are also attributes that from an accounting perspective represent a residual in the sense that they accommodate other non-interest bearing-assets and liabilities, that cannot be registered in the other attributes. The B82 was also the attribute where the greatest number of outliers detected for the boxplot. This is explained by the fact that this attribute has usually constant behaviour, mainly equal to zero, but when it changes tends to make it spontaneous and in an expressive way.

In the Tables 8, 9 and 10 is represented the number of outliers in an attribute for each CAE. As an example, in the Table 8 for companies with CAE 1 were identified 17 outliers in the attributes B05 and 42 outliers for companies with CAE 2. Consequently, it will be analysed the CAE which is more affected by the presence of outliers and the attributes that contributes more to this result.

Table 8: Number of outliers detected per attribute for each CAE with z-score method for companies with less than 20 observations, in absolute frequencies.

CAE	B05	B10	B15	B25	B30	B41	B51	B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	Total
1	17	27	39	37	22	91	76	38	26		38	1	47	28	10	26	26	2	551
2	42	38	97	77	87	122	122	67	56	15	46	9	74	42	24	39	46		1003
3	41	61	71	52	88	101	112	101	75	6	47	2	50	66	16	74	45	22	1030
4	18	26	66	23	61	80	78	43	40	8	39	2	73	64	16	41	48	2	728
5		1		8	7	2	4	1			1		1		1	1	1		28
6	11	38	69	35	60	84	49	54	46		26	6	40	46	6	48	29	4	651
7	29	38	63	51	61	120	76	63	72	6	39	7	74	40	13	54	51	4	861
8	45	40	86	44	114	103	103	41	36	5	39	8	63	80	34	52	91	11	995
9	11	34	46	33	50	61	38	36	23	1	18	5	43	11		47	13		470
10	29	34	60	42	87	88	67	42	26	6	42	4	48	68	2	16	26	4	691
11	17	46	58	37	55	69	70	53	42		26	4	28	37	11	30	41		624
12	34	46	82	54	74	105	87	81	89	1	41	2	29	41	7	48	36		857
13	24	73	14	123	50	108	84	33	25	1	30	9	59	25		29	55	3	745
14	6	20	28	21	41	62	65	44	51		32	8	57	8	6	31	18		498
18	45	88	53	66	64	130	132	60	66	11	36	6	102	68	24	62	40	2	1055
19	37	74	95	63	77	119	73	88	73	5	45	10	95	71	2	84	90	3	1104
20	9	14	24	29	13	30	22	14	10	3	18		24	9	1	6	4		230
21	36	153	49	139	71	150	176	58	67	1	80	7	104	53	57	111	22	5	1339
22	49	96	83	117	115	163	152	115	88	4	55	16	121	91	15	72	53	2	1407
23	35	73	77	113	92	149	103	81	77	2	43	9	97	49	10	66	24	8	1108
24	19	47	16	49	11	59	35	30	17		18	1	33	10	4	34	1		384
25	24	61	14	48	27	101	71	36	39	7	31	6	43	23	4	22	19	1	577
26	16	68	54	59	45	82	61	53	60		30	4	66	23	8	51	28	2	710
27	33	75	32	62	26	80	109	65	51	23	51	33	75	58	1	33	33	5	845
41	37	84	71	102	56	143	156	82	79		84	7	102	85	47	105	39	5	1284
42	18	81	33	64	51	82	102	49	45	6	31	3	64	46	23	46	14	5	763
43	39	39	68	66	75	80	78	53	32		35	4	81	45	20	54	18		787
51	21	39	71	52	69	93	93	78	69	15	56	1	73	32	7	62	41	6	878
52	60	109	227	151	214	274	223	156	155	15	72	9	180	117	44	103	125	23	2257
53	28	109	80	103	86	152	181	72	70	10	39	1	95	49	30	77	53	5	1240
60	58	51	71	76	97	142	136	85	96	12	59	17	89	56	16	85	32	4	1182
61	12	57	8	67	14	93	76	42	31	1	24	1	62	44	1	54	32	6	625
Total	900	1840	1905	2063	2060	3318	3010	1914	1732	164	1271	202	2192	1485	460	1663	1194	134	27507

Table 9: Number of outliers detected per attribute for each CAE with z-score method for companies with at least 20 observations, absolute frequencies.

CAE	B05	B10	B15	B25	B30	B41	B51	B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	Total
1	38	26	24	21	21	99	67	56	82	6	49	33	68	87	18	23	35	16	769
2	112	62	139	118	189	229	257	143	199	68	121	36	277	223	63	65	182	32	2515
3	160	61	83	59	120	157	187	137	130	54	91	16	181	204	49	128	161	26	2004
4	107	62	130	45	110	188	117	100	100	56	86	47	185	202	51	84	124	12	1806
5	2	1	7	3	8	5	8	7	6		1		6						54
6	135	73	188	47	120	170	189	126	217	34	155	27	200	209	46	45	196	10	2187
7	65	48	81	43	87	163	143	178	202	49	102	27	216	254	48	85	197	31	2019
8	76	65	71	54	125	189	112	118	124	45	75	12	174	161	27	82	176	19	1705
9	31	75	85	45	65	151	108	140	105	32	112	24	160	155	23	97	112	7	1527
10	45	27	39	34	63	83	110	88	101	16	60	13	64	77	34	34	88		976
11	66	41	80	42	97	158	166	148	159	39	98	33	113	86	18	49	109	30	1532
12	83	55	65	50	81	96	84	84	120	22	44	12	84	95	24	34	71	2	1106
13	85	95	88	199	133	169	141	103	71	15	188	24	186	97	26	51	147	21	1839
14	22	33	69	64	60	102	90	32	44	3	33	2	64	71	23	22	49	5	788
18	37	46	44	53	97	117	109	91	99	17	67	27	103	69	26	69	68	9	1148
19	51	54	112	71	124	115	107	150	139	17	83	17	163	159	3	111	154	4	1634
20	50	19	77	25	51	75	63	79	67	2	64	27	72	37	13	22	64	9	816
21	85	154	25	130	74	140	145	72	76		100	35	136	91	72	98	79	8	1520
22	83	97	134	92	138	128	130	107	96	17	94	36	178	88	15	110	87	7	1637
23	58	41	99	107	131	110	137	107	130	26	78	46	133	120	29	68	127	17	1564
24	44	29	26	18	23	56	46	41	47		39	1	53	52	12	11	37	5	540
25	23	37	26	37	29	103	84	59	47	56	69	21	91	102	7	26	69	2	888
26	23	56	64	63	60	130	78	53	69	27	57	9	101	90	50	40	35	1	1006
27	146	194	84	222	115	197	262	132	155	107	141	158	251	72	24	149	134	44	2587
41	21	42	39	56	64	98	119	27	45	8	39	14	57	61	50	60	36	33	869
42	119	59	85	76	120	76	77	64	45	38	68	67	150	111	47	49	142	43	1436
43	33	22	43	48	68	37	45	61	44	11	20	12	73	54	34	61	60	4	730
51	80	68	201	74	106	201	96	127	129	29	92	35	149	147	33	155	134	5	1861
52	263	154	464	171	380	562	497	469	488	103	295	101	476	423	113	273	407	43	5682
53	92	124	135	74	84	183	179	121	158	26	76	31	199	186	40	26	97	19	1850
60	84	158	85	107	136	191	168	200	149	32	92	28	229	152	10	150	159	13	2143
61	122	232	174	292	262	462	398	270	280	59	164	11	283	288	95	260	382	51	4085
Total	2441	2310	3066	2540	3341	4940	4519	3690	3923	1014	2853	982	4875	4223	1123	2537	3918	528	52823

Table 10: Number of outliers detected per attribute for each CAE with boxplot for companies with at least 20 observations, absolute frequencies.

CAE	B05	B10	B15	B25	B30	B41	B51	B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	Total
1	11	9	10	10	14	28	9	21	24	2	13	11	27	11	8	15	9	3	235
2	74	13	52	35	50	58	52	33	67	21	47	25	138	58	24	25	86	28	886
3	79	17	19	10	38	40	33	22	39	16	36	10	111	41	29	33	32	3	608
4	45	7	41	16	34	42	26	16	18	14	28	18	86	47	17	13	28	5	501
5		1	2	1	2		4				1		6						17
6	59	25	41	7	25	37	34	35	39	9	48	12	62	36	17	17	41	12	556
7	50	8	37	22	22	35	18	32	62	17	29	25	100	60	18	23	51	11	620
8	51	26	20	11	29	37	29	36	30	6	29	13	67	28	12	20	36	3	483
9	19	9	19	16	23	36	22	41	39	5	28	10	73	26	9	27	53	7	462
10	21	14	8	8	27	31	8	19	17	9	21	7	38	20	11	20	19		298
11	34	18	22	8	29	29	22	37	51	10	51	20	52	24	10	12	32	6	467
12	36	22	17	8	16	29	23	9	22	12	17	10	37	15	14	17	33	2	339
13	41	26	29	58	46	20	34	20	27	11	18	9	89	53		21	43	18	563
14	15	7	25	8	27	18	24	20	22	3	4	2	37	18	7	22	23	5	287
18	30	15	31	24	28	37	13	16	29	6	37	16	42	13	10	24	14	4	389
19	43	16	24	10	33	20	10	42	41	6	22	12	52	29	3	13	45		421
20	19	16	17	4	11	21	19	22	10	2	14	8	41	28	3	3	12	6	256
21	16	21	7	44	34	22	33	24	31		19	18	42	23	29	36	35	4	438
22	26	29	41	32	54	29	32	15	25	2	21	26	73	28	3	25	33	3	497
23	26	11	49	20	51	23	25	25	39	6	26	21	81	32	7	11	48	11	512
24	5	12	10	3	8	13	7	17	15		7		17	17	7	5	13	5	161
25	18	12	15	11	14	14	13	17	17	7	14	7	26	11	10	10	16	2	234
26	16	24	33	12	30	37	17	17	13		16	9	49	29	9	26	11	1	349
27	62	59	39	86	52	54	54	33	52	16	57	56	75	11	7	14	49	18	794
41	8	9	15	23	21	19	19	16	15	2	14	2	31	14	11	17	18	11	265
42	40	14	16	22	27	22	9	17	19	4	37	31	30	19	19	10	23	34	393
43	13	12	14	3	10	5	10	18	14	3	18	12	18	10	14	5	10	4	193
51	39	20	51	23	30	30	14	45	49	17	26	12	56	29	17	41	27	5	531
52	148	36	132	73	116	183	106	163	180	20	124	59	215	107	31	62	137	28	1920
53	41	64	43	21	18	44	30	36	47	7	19	22	77	48	16	14	19	5	571
60	45	53	41	42	49	31	18	46	51	10	28	15	60	31	13	24	42	3	602
61	59	69	60	91	90	96	73	99	112	18	24	7	153	71	14	53	82	34	1205
Total	1189	694	980	762	1058	1140	840	1009	1216	261	893	505	2061	987	399	658	1120	281	16053

The results shown in the Tables 8, 9, 10 are in absolute frequency. So, the number of outliers detected is strongly affected by the number of observations per CAE, which is not homogenous. Thus, to facilitate analysis in the Annex E are presented the results in relative frequencies of the outliers detected for the total number of observations per CAE.

In the Table 9 is presented the number of outliers detected in an attribute for each CAE with z-score method for companies with less than 20 observations. In this implementation the results at the CAE level are almost homogeneous, although it is verified some CAE´s with a higher percentage than the remaining ones, these are not significant enough to justify a more detailed analysis. However, in Table 21 (Annex E) for CAE 5 in the attributes B25 and B30 an outlier percentage of 16.67% and 14.58% respectively was verified. This high percentage is explained little representativeness of this CAE 5 in the sample. For companies with less than 20 observations this CAE is only represented by 18 companies, causing outlier detected to have more impact in terms of percentage.

The CAE analysis for companies with at least 20 observations for the z-score and boxplot presented in the tables 9 and 10 reflect similar conclusions. The CAE with the most relevant percentage of outlier is the 27 (Tables 22 and 23, Annex E), which refers to companies related to the manufacture of electrical products such as engines, energy generators, electronic appliances and others. As they are production companies, they tend to stock great amount of supplies in order to perform more profitable deals. Thus, it is natural that the account related to the suppliers (B25) suffer from great fluctuations. Other attributes that explain the high percentage of outliers in this CAE are B76 and B78. As explained above, these are related to form of funding normally associated with large companies. In this CAE, the sample is mainly represented by small and medium-sized companies (Table 24, Annex F), so it is not usual to use this method types of funding. Nonetheless, when they do so, they tend to be short term operations.

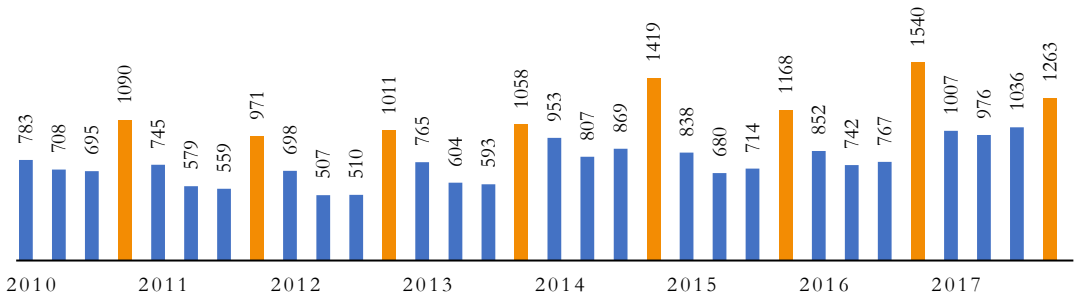

Figure 15 - Number of outliers detected quarterly from 2010 to 2017, with z-score method for companies with less than 20 observations.

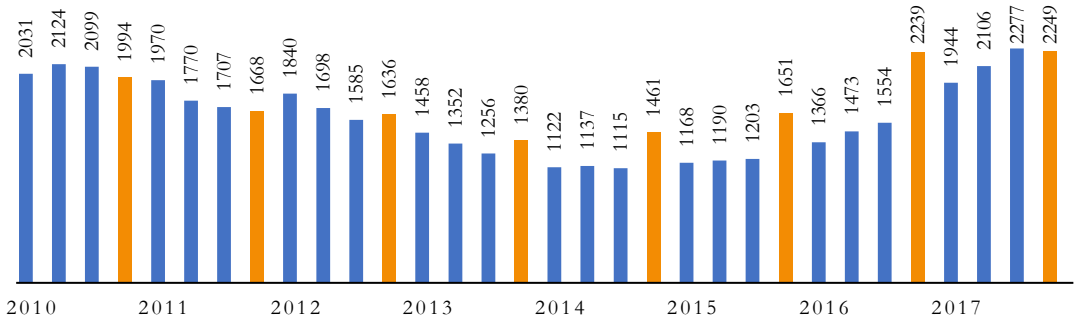

Figure 16 - Number of outliers detected quarterly from 2010 to 2017, with z-score method for companies with at least 20 observations.

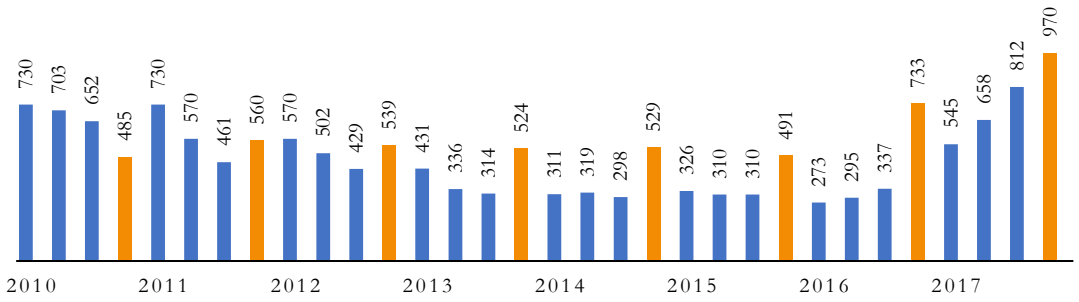

Figure 17 - Number of outliers detected quarterly from 2010 to 2017, with z-boxplot method for companies with at least 20 observations.

The Figures 15, 16 and 17 is presented the outliers detected on a quarterly basis. The orange bars represent the fourth quarter of each year and is typically where the higher number of outliers are detected. This is related with a set of procedures consistent with the accounts closing. These have an impact on impairment, depreciation, amortization, taxes and other accounts, which may cause some seasonal impact in the attributes.

In the methods applied to companies with at least 20 observations it is possible to verify that in 2013 there was a decrease in the number of outliers detected. In this year was when the extrapolation factor began to be imputed to the companies, leading to a change in the

selection process of companies' subject to quality control processes. Previously, the selection was based on large differences, and consequently very focused on large companies. The extrapolation factor gave small and medium companies more statistical relevance. Causing them to be more susceptible to quality control and consequently to an overall higher quality of data. The increase in the year 2017 is explained by the fact that at the time the dataset was provided, this period had been subject to a limited quality control process.

In the method applied to companies with less than 20 observations the impact of the introduction of extrapolation factor is not so significative. These are companies that enter and leave the sample and are most small and medium companies. In this way, as companies are not always in the sample, the improvement of quality control is not so visible because their presence in the sample is circumstantial. The number of outliers detected in the fourth quarter is even more significative for these companies, which is also explained by the procedures related to the accounts closing, as large companies make these adjustments more frequently, for example on a quarterly basis.

4.2 Multivariate Results

At the Multivariate level will be presented the results of the LOF and DBSCAN method. The multivariate method is applied at the observations level, making it impossible to specify which attributes may contribute in their detection. Thus, the analysis performed at this level will only consider the number of outliers detected per period and by CAE. In order to avoid the repetition of conclusions, the presentation and discussion of results of the two methods will be simultaneous, due to their similarity of results.

As previously stated, the definition of LOF outlier score is based on a visual analysis. In the Figures 18 and 19, the LOF scores are sorted in an ascending order for *MinPts* equal to 3 and 19. Then, the separation of the scores was made at the end of the "valley". The criterion for selecting the score is based on "significant" variations on the calculated scores, which introduces a degree of subjectivity in the results. Another alternative could be the application of a method of detection of extreme values like Boxplot to the obtained results. In Table 11 is represented the different alternatives and the impact in the number of the outliers detected.

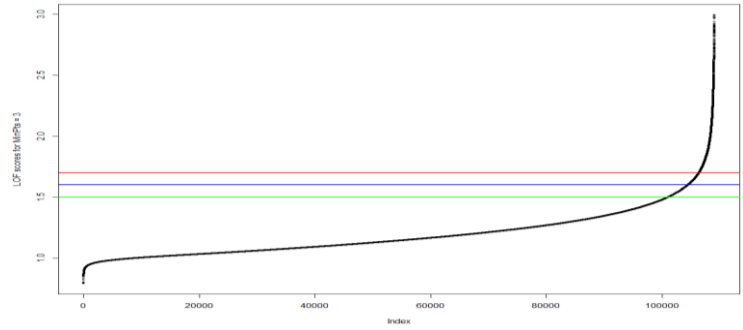

Figure 18 - Ordered Local outlier scores for *MinPts*=3 separated at 1.5 (green), 1.6 (blue) and 1.7 (red) scores.

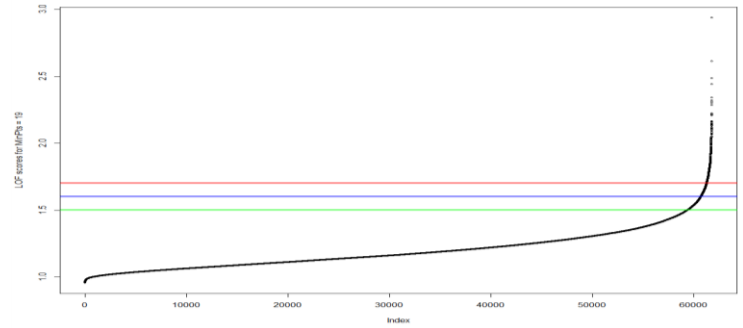

Figure 19 - Ordered Local outlier scores for *MinPts*=19 separated at 1.5 (green), 1.6 (blue) and 1.7 (red) scores.

Table 11: Number of detected outliers for the different threshold separation of Local outlier scores and its representativeness in the dataset.

Threshold	MinPts = 3	%	MinPts = 19	%
1.5	8131	7.44	2390	3.87
1.6	4737	4.33	1094	1.77
1.7	2831	2.59	513	0.83

Therefore, will be considered as outlier for the LOF with *MinPts* of 3 a score equal or higher than 1.6. And for a LOF with *MinPts* of 19 a score equal or higher than 1.5. The choice considers the visual analysis of the graphs above but also because a percentage around 4% of outliers should be acceptable. Nonetheless, if applied the method with a k = 1.5 ("moderate outlier") to the LOF scores with *MinPts* of 3, the upper whisker will be set at 1.62 and the outlier percentage at 3.84%. For the LOF scores with *MinPts* of 19, the upper whisker will be set at 1.54 and the outlier percentage at 2.86%, which means that the definition of the outlier score by both methods is close.

Table 12: Descriptive statistics of LOF scores with *MinPts* of 3 and 19.

MinPts	Min.	1º quartile	Median	Mean	3º quartile	Max.	Std.
3	0.8	1.1	1.1	9.6	1.3	433027.1	1765.63
19	0.96	1.09	1.16	1.2	1.27	3.11	0.15

The descriptive statistics table for the two implementations of the LOF shows that the implementation with *MinPts* of 19 for companies with at least 20 observations has more homogeneous results, because the local density is calculated considering more neighborhood points. It is also emphasized that in the implementation with *MinPts* of 3 for companies with less than 20 observations there are some observations with very high scores. Generally, are observations close to other observations that shared the same or a very close spatial coordinates with other observations. So, are observations with low reachability distance compared to their neighbors receiving a high LOF score.

The DBSCAN method unlike LOF has its output in a discrete form, assigning a cluster to each observation. Thus, to the outlier observation are assigned the cluster zero. In order not to cause confusion to the readers it is remembered that *MinPts* has a different definition in this method. So, *MinPts* is the least number of observations required to form a dense region. In Table 13 is represented the cluster results for the two different implementations.

Table 13: Clusters obtained in the implementation of this method for *MinPts* equal to 4 and 20 are represented in the following table.

Cluster	0	1	2	3	4	5
MinPts = 4	3228	106029	8	4	4	4
MinPts = 20	2118	59666	-	-	-	-

For both implementations the DBSCAN method basically agglomerates all companies that are not outlier in the same cluster. Although it is not possible to visualize the definition of the clusters in an 18-dimensional plot, it is known that this method can handle arbitrarily shaped clusters. This result can be related to the choice of distance function. Since for the Euclidean, distance the increase of the dimensional space causes the data to become sparse, and the distance between all pairs of observations to become less meaningful (Hinneburg et al., 2000). So, it is possible the existence of multiple border points connecting all the different density regions leading to a creating of a single connected component. Nonetheless, for the *MinPts* set to 4 applied for companies with at least 4 observations were detected a total of 3.288 outliers, representing around 2.95% of the total observations. For the *MinPts* set to 20 applied for companies with at least 20 observations were detected a total of 2.118 outliers, representing around 3.43% of the total observations. In the Table 14 is represented the number of outliers detected by CAE for the different implementations of multivariate methods.

Table 14: Absolute and relative frequencies of the number of outliers detected for each CAE with LOF and DBSCAN.

	Absolute Frequencies				Relative Frequencies (%)			
	LOF		DBSCAN		LOF		DBSCAN	
CAE	MPts = 3	MPts = 19	MPts = 4	MPts = 20	MPts = 3	MPts = 19	MPts = 4	MPts = 20
1	94	52	65	52	4.84	5.24	3.35	5.24
2	258	155	200	127	5.48	5.04	4.25	4.13
3	216	119	137	95	4.71	4.44	2.99	3.54
4	179	99	111	73	5.14	4.57	3.19	3.37
5	4	4	3	3	3.20	5.19	2.40	3.90
6	142	110	123	103	4.14	4.64	3.59	4.34
7	210	116	159	91	5.06	4.46	3.83	3.50
8	170	87	131	55	4.28	4.02	3.29	2.54
9	97	59	69	59	3.67	3.29	2.61	3.29
10	99	49	63	44	3.98	3.48	2.53	3.13
11	113	58	73	50	3.49	2.61	2.25	2.25
12	112	53	93	56	3.75	3.54	3.12	3.74
13	139	82	64	61	4.20	3.89	1.93	2.90
14	90	54	59	40	4.48	4.74	2.94	3.51
18	146	47	87	43	4.23	3.35	2.52	3.06
19	104	49	94	52	3.22	3.05	2.91	3.23
20	31	25	38	28	2.70	3.49	3.30	3.91
21	244	87	118	72	6.40	5.49	3.10	4.55
22	196	63	122	63	4.80	3.55	2.99	3.55
23	149	73	89	63	4.11	4.38	2.46	3.78
24	60	24	32	18	3.73	3.34	1.99	2.50
25	69	42	59	46	3.07	3.99	2.62	4.37
26	99	35	65	36	3.91	2.54	2.57	2.62
27	241	144	143	106	7.86	7.63	4.66	5.62
41	156	47	106	40	5.18	4.73	3.52	4.02
42	108	64	71	49	4.08	4.75	2.68	3.64
43	118	37	60	30	4.81	4.04	2.45	3.28
51	162	71	129	81	4.52	3.58	3.60	4.08
52	416	207	307	221	4.03	3.15	2.97	3.36
53	159	62	130	71	3.65	3.24	2.98	3.72
60	163	74	128	78	3.42	2.77	2.69	2.91
61	193	142	100	112	3.09	2.68	1.60	2.11
Total	4737	2390	3228	2118	4.33	3.87	2.95	3.43

Similarly to what was verified for univariate implementations, is in CAE 27 where the largest number of cases were detected. This demonstrates consistency of the results for the different methods and the effect of univariate outliers at the multivariate level. In the Figures below it is represented the number of outliers detected on a quarterly basis for the different implementations.

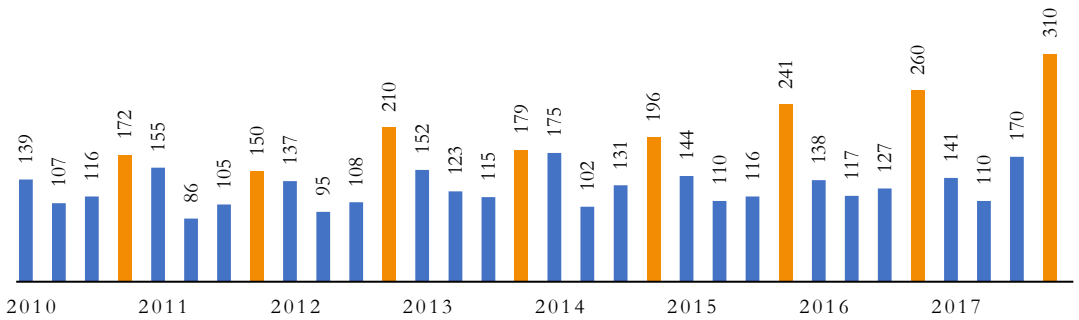

Figure 20 - Number of outliers detected quarterly from 2010 to 2017, with LOF method for MinPts = 3.

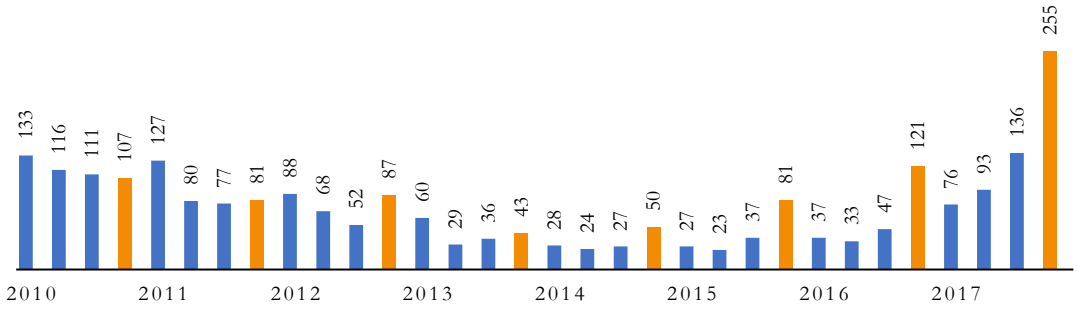

Figure 21 - Number of outliers detected quarterly from 2010 to 2017, with LOF method for *MinPts* = 19.

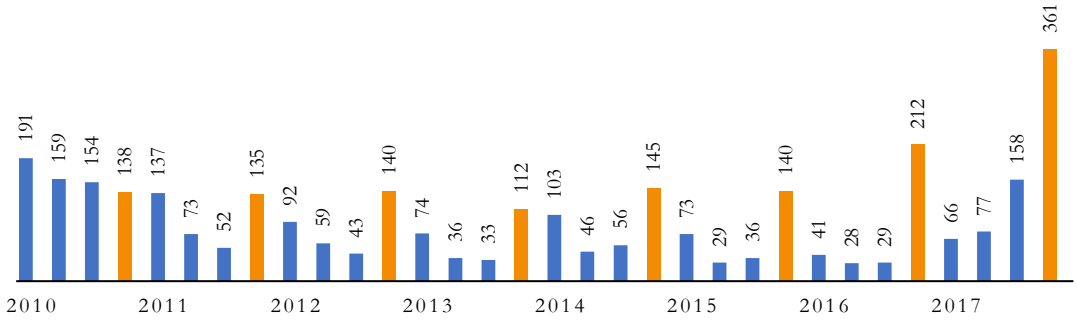

Figure 22 – Number of outliers detected quarterly from 2010 to 2017, with DBSCAN method for *MinPts* = 4.

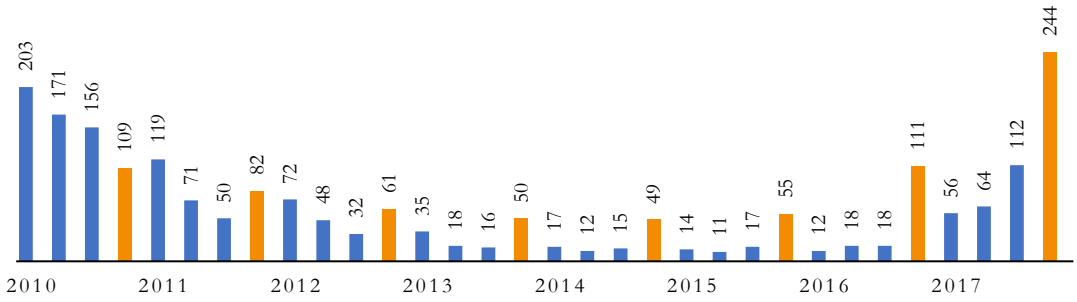

Figure 23 - Number of outliers detected quarterly from 2010 to 2017, with DBSCAN method for *MinPts* = 20.

Similarly, to the CAE analysis, for the outliers detected per quarter period demonstrated in the graphs above, the results are also consistent with those obtained for the univariate methods. To summarize, in the implementation for companies with more than 20 observations it is noticeable the effects of the improving in the quality control processes carried out from 2013 onwards and in a greater number of outliers detected for the fourth quarter due to a set of procedures consistent with the accounts closing.

4.3 Comparison of results

During the presentation of results some comparisons have already been made between the results obtained for the different implementations. In this section will be presented the general comparison of the results and analyzed the overlap of the outliers detected in the different implementations.

In the Table 15 it is presented the results of the univariate methods. As expected, the boxplot detects less outliers than z-score. It is also perceptible that the percentage of attributes where outliers were detected is high, which is a result of the variability presented in the data. However, when considered the total number of observations the results are acceptable. It is also emphasized that for companies with more than 20 observations, around 82% of the outliers detected by the boxplot are also detected by the z-score. So, the results presented for both methods are consistent.

Table 15: Comparison of univariate methods.

	Z-score <20	Z-score >20	Boxplot	Atr. Out (%)	Obs. Out (%)
Z-score <20	27507	0	0	13.34	3.22
Z-score >20		52823	13055	30.76	4.75
Boxplot			16053	18.30	1.49

In the Figure 24 is presented a Venn diagram for the multivariate methods, which makes possible to visualize the overlap of the detected outliers for the different implementations. In the diagram and Table 16, the results are presented considering the entire dataset. However, in Table 25 (Annex G) is specified the outliers detected for the companies with at least 20 observations. The main conclusion of this comparison of results is that 464 of the outliers detected are common to all implementations. In addition, the method with less shared outliers detected was the LOF with 3 *MinPts*. Both LOF implementations only shared 809 outliers detected, which represent around 35% of the cases. This proves the influence of the parameterization in this method, as well as the influence of the number of *MinPts* in the local density definition. On the other hand, for companies with at least 20 observations all the outliers detected by DBSCAN with 4 *MinPts* were also detected by the DBSCAN with 20 *MinPts*. Although the implementation with *MinPts* 20 detects more 501 outliers, the influence of this parameter is not as significant as in the LOF. When comparing the two methods in general, it is noticeable that these have a low percentage of outliers shared. This can be explained by the different definitions of local density in the two methods. Nevertheless, it is emphasized that most of the outliers detected by the DBSCAN are also

associated with a high value of LOF score, which demonstrates the influence of the definition of the score in the method and in our analysis.

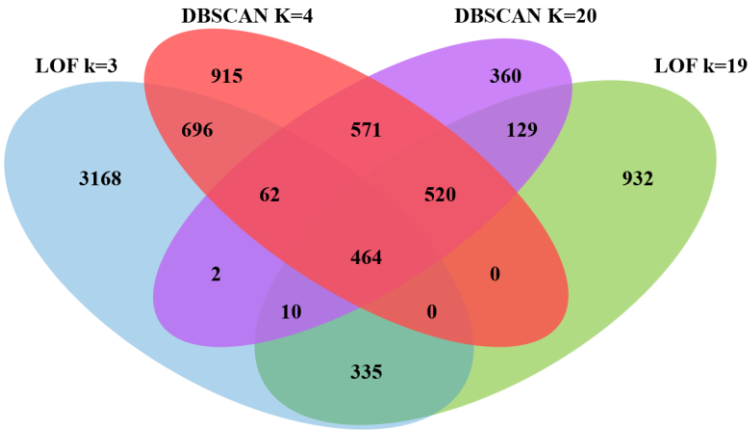

Figure 24 - Venn diagram of the outlier detected in the multivariate methods, K represents the applied *MinPts* parameter.

In the Table 16, a summary of detected outliers of the different methods for the entire dataset is presented. In order to make the comparison between univariate and multivariate methods, it was considered as outlier an observation in the multivariate space that contains univariate outliers.

Table 16: Comparison of the detected outliers for the different methods.

	LOF K=3	LOF k=19	DBSCAN k=4	DBSCAN K=20	Boxplot	Z-score
LOF K=3	4737					
LOF K=19	809	2390				
DBSCAN K=4	1222	984	3228			
DBSCAN K=20	538	1123	1617	2118		
Boxplot	869	1535	1480	1906	12458	
Z-score	2872	2080	3012	2049	11485	48406
Out (%)	4.33	3.87	2.95	3.43	20.16	44.32

When looking to the results of both DBSCAN implementation it is perceptible that a very high percentage of the flagged observations are common outliers to the boxplot. Although in LOF implementations this percentage it is not high it is still significant. Therefore, it is possible to assume a correlation between the detection of extreme values in univariate space and multivariate outliers. Since the presence of extreme values directly influences the calculation of the Euclidean distance of an observation towards its neighbors, and consequently the definition of its local density.

Chapter 5 - Conclusion and Future Work

The main objective of this dissertation was to identify a group of situations susceptible to be delivered for manual analysis by BDP technicians, optimizing the human resources in the quality control process of the ITENF database, through the implementation of outlier detection methods.

Univariate methods, such as z-score and Boxplot, and multivariate methods, such as LOF and DBSCAN were applied. In the analysis of univariate methods, the z-score was the method with the highest percentage of outliers detected. This method tends to struggle for company attributes with high variability and different distribution, meaning that there could be a large number of outliers falsely detected. Nonetheless, this method performs better for companies with less than 20 observations, identifying mostly isolated extreme values. As expected, attributes with more observations also have more volatility, since they reflect the evolution of the corporate behaviour over a longer period.

The boxplot was only implemented for companies with at least 20 observations, to ensure that the definition of quartiles has some meaning. Compared to the z-score, the number of outliers detected is less than half. Since, this method is more robust and tends to deal better with high variability, it should flag a lower percentage of false outliers. However, in both univariate methods were found attributes with a large number of detected outliers, which are usually associated with arbitrary behaviour and with the temporal component presented in the data. This means that when a temporary or permanent shift of the attribute values is verified multiple outliers tend to be flagged.

In the analysis of outliers detected by attribute through the univariate methods, it is verified that the attributes with less detected outliers were the B70, B78, C60 and C90. These attributes tend to be more stable or non-existent in several companies, since are usually related with obtained funding and accounting operations that are generally practiced by large companies. In other hand, the attributes with more detected outliers were the B41, B51 and B82. The B70 and B78 are attributes directly affected by the volatility, seasonality of the company´s activity and from an accounting perspective represent a residual. The B82 is explained by the attribute usual constant behaviour, that when varies tend to make it spontaneous and expressively.

For multivariate methods, the LOF tends to find more outliers than the DBSCAN. The definition of LOF score has a great impact on the results obtained. The outliers detected by DBSCAN usually are associated with high LOF scores. So, a different choice on the LOF score would result in a greater overlap of detected common cases. In terms of parametrization the LOF is more volatile than the DBSCAN, since the in interception of results for companies with more than 20 observations, all the outlier detected with *MinPts* of 4 were also detected for *MinPts* equal to 20. While in the LOF the interception for different parameterizations is only around 35%.

When analysing the number of detected outliers for each CAE and by quarter, it is perceptible that the conclusions drawn are consistent for both univariate and multivariate methods. The CAE with the most relevant percentage of outliers is the 27, which represents the companies related to the manufacture of electrical products. These companies tend to suffer from great fluctuations with accounts related to the suppliers, since they usually stock great amounts of supplies to perform more profitable deals. For companies with at least 20 observations it is noticeable the effects of the improvements in the quality control processes carried out from 2013 onwards. It is also noticeable that there are a higher number of outliers detected overall in the fourth quarter due to a set of procedures consistent with the accounts closing.

Considering the characteristics of the BDP dataset, we should resort to several more methods to optimize the process of selection of companies. At the univariate level, the z-score should be applied for companies with less than 20 observations and the boxplot to the remaining. So, it would be possible to discover great variations in the dataset and where are located. Differently, multivariate methods will identify a set of observations where the combined behavior of theirs attributes raise interest for manual analysis. The LOF is the more suitable method for this application. Although its results are more volatile when considering its parameterization, the output is in the form of a score which allows the experts more freedom to define the number of outliers that should be analyzed. While in DBSCAN the output is in discrete labels, and the threshold is defined during the implementation, making it more difficult to modify the selection process of the method.

In terms of future work, time series methods will become more suitable with the increase of observations for larger companies, as these consider the temporal component of the data. For instance, Chen & Liu (1993) developed a time series method that automatically detects

outliers, while simultaneously estimating the parameters of a fitted model and the effect of the outliers on the model. This is a univariate method that can recognize four types of outlier such as addictive outlier (AO), level shift (LS), temporary change (TC) and innovative outlier (IO). Another possibility is the implementation of the "Seasonal and Trend decomposition using Loess" also known as STL, which decompose the time series into trend, seasonal and remainder components (Cleveland et al., 1990). The major advantages of STL is that can deal with missing values. Researchers will be able to apply these methods in the upcoming years.

References

[1] "What is R?", https://www.r-project.org/about.html, accessed in 15 August 2018.

[2] "Modified Z score", https://www.ibm.com/support/knowledgecenter/en/SS4QC9/com.ibm.solutions.wa_an_overview.2.0.0.doc/modified_z.html, accessed in 23 March 2018.

[3] "Mean and median absolute deviation", https://www.ibm.com/support/knowledgecenter/no/SSWLVY_1.0.0/com.ibm.spss.analyticcatalyst.help/analytic_catalyst/meanmedianabsolutedeviation.html, accessed in 23 March 2018.

Abe, N., Zadrozny, B., & Langford, J. (2006, August). Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 504-509). ACM.

Acuna, E., & Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez*.

Acuna, E. & Rodriguez, C. (2005). dprep package. R package version 3.0.2. https://CRAN.R-project.org/package=dprep.

Aggarwal, C. C. (2017). Outlier analysis. Springer International Publishing.

Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, *55*, 278-288.

Angiulli, F., & Pizzuti, C. (2002, August). Fast outlier detection in high dimensional spaces. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 15-27). Springer, Berlin, Heidelberg.

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999, June). OPTICS: ordering points to identify the clustering structure. In *ACM Sigmod record* (Vol. 28, No. 2, pp. 49-60). ACM.

Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data (Probability & Mathematical Statistics).

Barkan, O., & Averbuch, A. Robust Subspace Mixture Models for Anomaly Detection in High Dimensions. *IEEE Transactions on Journal NAME, Manuscript ID*.

Ben-Gal, I. (2005). Outlier detection. *Data mining and knowledge discovery handbook*, 131-146.

Battipaglia, P., Bruno, G., Farabullini, F., Mahlknecht, C. (2004). Selective editing to improve the quality of banking statistics. *Supplement to the statistical bulletin* (Vol. 14, No. 29). Banca D´ITALIA.

Bay, S., Kumaraswamy, K., Anderle, M. G., Kumar, R., & Steier, D. M. (2006, December). Large scale detection of irregularities in accounting data. In *Data Mining, 2006. ICDM'06. Sixth International Conference on* (pp. 75-86). IEEE.

Boriah, S., Chandola, V., & Kumar, V. (2008, April). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining* (pp. 243-254). Society for Industrial and Applied Mathematics.

Bramati, M. C., & Croux, C. (2007). Robust estimators for the fixed effects panel data model. *The econometrics journal*, *10*(3), 521-540.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In *ACM sigmod record* (Vol. 29, No. 2, pp. 93-104). ACM.

Bowley, A. L. (1920). *The change in the distribution of the national income, 1880-1913*. Oxford: Clarendon Press.

Campello, R. J., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *10*(1), 5.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 15.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, *6*(1), 1-6.

Chen, C. and Liu, Lon-Mu (1993). "Joint Estimation of Model Parameters and Outlier Effects in Time series". *Journal of the American Statistical Association*, 88(421), pp. 284-297.

Cleveland, R. B., W. S. Cleveland, J. E. McRae and I. J. Terpenning (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* 6(1), 3-73.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15-18.

Crone, B. A., Midgley, M. S., & Sanders, J. (2012). Lake Dwellings after Robert Munro. Proceedings from the Munro International Seminar: The Lake Dwellings of Europe, 22nd and 23rd October 2010, University of Edinburgh.

Dali, L., Bentajer, A., Abdelmajid, E., Abouelmehdi, K., Elsayed, H., Fatiha, E., & Abderahim, B. (2015, March). A survey of intrusion detection system. In *Web Applications and Networking (WSWAN), 2015 2nd World Symposium on* (pp. 1-6). IEEE.

Daneshpazhouh, A., & Sami, A. (2015). Semi-supervised outlier detection with only positive and unlabeled data based on fuzzy clustering. *International Journal on Artificial Intelligence Tools*, *24*(03), 1550003.

Davidson, I. (2007). Anomaly detection, explanation and visualization. *SGI Technical Report*.

Deneshkumar, V., Senthamaraikannan, K., & Manikandan, M. (2014). Identification of outliers in medical diagnostic system using data mining techniques. *International Journal of Statistics and Applications*, *4*(6), 241-248.

Deshpande, M., & Karypis, G. (2002, May). Evaluation of techniques for classifying biological sequences. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 417-431). Springer, Berlin, Heidelberg.

Diehl, C. and Hampshire, J. (2002). Real-time object classification and novelty detection for collaborative video surveillance. *In Proceedings of the IEEE International Joint Conference on Neural Networks*. IEEE.

Ding, Z., & Fei, M. (2013). An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes*, *46*(20), 12-17.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics*, *21*(4), 488-506.

Do, K., Tran, T., Phung, D., & Venkatesh, S. (2016). Outlier detection on mixed-type data: An energy-based approach. *In Advanced Data Mining and Applications: 12th International Conference, ADMA 2016, Gold Coast, QLD, Australia, December 12-15, 2016, Proceedings* 12 (pp. 111-125). Springer International Publishing.

Edgeworth, F. Y. (1887). Xli. on discordant observations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *23*(143), 364-375.

Elkan, C., & Noto, K. (2008, August). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 213-220). ACM.

Ertöz, L., Steinbach, M., & Kumar, V. (2003, May). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 2003 SIAM international conference on data mining* (pp. 47-58). Society for Industrial and Applied Mathematics.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Flach, M., Gans, F., Brenning, A., Denzler, J., Reichstein, M., Rodner, E., ... & Mahecha, M. D. (2017). Multivariate anomaly detection for Earth observations: a comparison of algorithms and feature extraction techniques. *Earth System Dynamics*, *8*(3), 677.

Finch, W. H. (2012). Distribution of variables by method of outlier detection. *Frontiers in psychology*, *3*.

Gama, J., Carvalho, A. C. P. D. L., Faceli, K., Lorena, A. C., & Oliveira, M. (2015). Extração de conhecimento de dados: data mining.

Gao, J., Cheng, H.B., Tan P.N. (2006). Semi-supervised outlier detection, in: *Proceedings of the ACMSymposium on Applied Computing, vol. 1, ACM Press, Dijon, France*, 2006, pp. 635–636.

Ghosh-Dastidar, B., & Schafer, J. L. (2006). Outlier detection and editing procedures for continuous multivariate data. *Journal of Official Statistics*, *22*(3), 487.

Goktug, A. N., Chai, S. C., & Chen, T. (2013). Data analysis approaches in high throughput screening. In *Drug Discovery*. Intech.

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, *11*(4), e0152173.

Grubbs, F. E. (1969). Procedures for detecting outlying xobservations in samples. *Technometrics*, *11*(1), 1-21.

Guha, S., Rastogi, R., & Shim, K. (1998, June). CURE: an efficient clustering algorithm for large databases. In *ACM Sigmod Record* (Vol. 27, No. 2, pp. 73-84). ACM.

Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *26*(9), 2250-2267.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Hahsler, M. & Piekenbrock, M. (2017). dbscan package. R package version 1.1.1. https://CRAN.R-project.org/package=dbscan.

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, *24*(9-10), 1641-1650.

Hinneburg, A., Aggarwal, C. C., & Keim, D. A. (2000). What is the nearest neighbor in high dimensional spaces?. In *26th Internat. Conference on Very Large Databases* (pp. 506-515).

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*(2), 85-126.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers* (Vol. 16). Milwaukee, WI: ASQC Quality Press.

Jin, W., Tung, A. K., Han, J., & Wang, W. (2006, April). Ranking outliers using symmetric neighborhood relationship. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 577-593). Springer, Berlin, Heidelberg.

Joshi, M. V., Agarwal, R. C., & Kumar, V. (2001). Mining needle in a haystack: classifying rare classes via two-phase rule induction. *ACM SIGMOD Record*, *30*(2), 91-102.

Joshi, M. V., Agarwal, R. C., & Kumar, V. (2002, July). Predicting rare classes: Can boosting make any weak learner strong?. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 297-306). ACM.

Jolliffe, I (2002). Principal component analysis. *Wiley Online Library*.

Rousseeuw, P. J., & Kaufman, L. (1990). Finding groups in data. *Series in Probability & Mathematical Statistics* (Vol. 34, pp. 111-112.).

Khan, S. S., & Madden, M. G. (2009, August). A survey of recent trends in one class classification. In *Irish conference on artificial intelligence and cognitive science* (pp. 188-197). Springer, Berlin, Heidelberg.

Knorr, E. M., & Ng, R. T. (1998, August). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the international conference on very large data bases* (pp. 392-403).

Knorr, E. M., & Ng, R. T. (1999, September). Finding intensional knowledge of distance-based outliers. In *VLDB* (Vol. 99, pp. 211-222).

Kozak, M. (2009). Analyzing one-way experiments: a piece of cake of a pain in the neck?. *Scientia Agricola*, *66*(4), 556-562.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2009, November). LoOP: local outlier probabilities. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1649-1652). ACM.

Kruskal, W. (1988). Miracles and statistics: The casual assumption of independence. *Journal of the American Statistical Association*, *83*(404), 929-940.

Latecki, L. J., Lazarevic, A., & Pokrajac, D. (2007, July). Outlier detection with kernel density functions. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition* (pp. 61-75). Springer, Berlin, Heidelberg.

Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., & Kavsek, B. (2000, August). Informal identification of outliers in medical data. In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology* (Vol. 1, pp. 20-24).

Li, X., & Han, J. (2007, September). Mining approximate top-k subspace anomalies in multi-dimensional time-series data. In *Proceedings of the 33rd international conference on Very large data bases* (pp. 447-458). VLDB Endowment.

Liu, R. M., Babanajad, S. K., Taylor, T., & Ansari, F. (2015). Experimental study on structural defect detection by monitoring distributed dynamic strain. *Smart Materials and Structures*, *24*(11), 115038.

Lu, F. (2007, September). Uncovering fraud in direct marketing data with a fraud auditing case builder. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 540-547). Springer, Berlin, Heidelberg.

Markou, M., & Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal processing*, *83*(12), 2481-2497.

Murphy, M. K. (1998). Consensus development methods, and their use in clinical guideline development. *Health technology assessment*, *2*(3), 1-88.

Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering*, *14*(5), 1003-1016.

Omar, S., Ngadi, A., & Jebur, H. H. (2013). Machine learning techniques for anomaly detection: an overview. *International Journal of Computer Applications*, *79*(2).

Pahuja, D., & Yadav, R. (2013). Outlier detection for different applications: Review. *International Journal of Engineering Research & Technology (IJERT)*, *2*.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., & Faloutsos, C. (2003, March). Loci: Fast outlier detection using the local correlation integral. In *Data Engineering, 2003. Proceedings. 19th International Conference on* (pp. 315-326). IEEE.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000, May). Efficient algorithms for mining outliers from large data sets. In *ACM Sigmod Record* (Vol. 29, No. 2, pp. 427-438). ACM.

Rorabacher, D. B. (1991). Statistical treatment for rejection of deviant values: critical values of Dixon's" Q" parameter and related subrange ratios at the 95% confidence level. *Analytical Chemistry*, *63*(2), 139-146.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, *88*(424), 1273-1283.

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589). John wiley & sons.

Sabau, A. S. (2012). Survey of clustering based financial fraud detection research. *Informatica Economica*, *16*(1), 110.

Saini, D. K., Ahir, D., & Ganatra, A. (2016). Techniques and challenges in building intelligent systems: anomaly detection in camera surveillance. In *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2* (pp. 11-21). Springer International Publishing.

Schreyer, M., Sattarov, T., Borth, D., Dengel, A., & Reimer, B. (2017). Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks. *arXiv preprint arXiv:1709.05254*.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), 19.

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets* (Doctoral dissertation, University of Pittsburgh).

Sharma, A., & Panigrahi, P. K. (2013). A review of financial accounting fraud detection based on data mining techniques. *arXiv preprint arXiv:1309.3944*.

Shevlyakov, G., Andrea, K., Choudur, L., Smirnov, P., Ulanov, A., & Vassilieva, N. (2013, May). Robust versions of the Tukey boxplot with their application to detection of outliers. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 6506-6510). IEEE.

Shiffler, R. E. (1988). Maximum Z scores and outliers. *The American Statistician*, *42*(1), 79-80.

Singh, K., & Upadhyaya, S. (2012). Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, *9*(1), 307.

Song, X., Wu, M., Jermaine, C., & Ranka, S. (2007). Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, *19*(5).

Szummer, M. O. (2002). *Learning from partially labeled data*(Doctoral dissertation, Massachusetts Institute of Technology).

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Introduction to Data Mining. *Addison-Wesley*.

Tang, J., Chen, Z., Fu, A. W. C., & Cheung, D. W. (2002, May). Enhancing effectiveness of outlier detections for low density patterns. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 535-548). Springer, Berlin, Heidelberg.

Tiao, G. C. (1985). 3 Autoregressive moving average models, intervention problems and outlier detection in time series. *Handbook of statistics*, *5*, 85-118.

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach.

Tukey, J. W. (1977). Exploratory data analysis.

Tsay, R. S., Peña, D., & Pankratz, A. E. (2000). Outliers in multivariate time series. *Biometrika*, *87*(4), 789-804.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., & Yin, K. (2003). A review of process fault detection and diagnosis: Part III: Process history based methods. *Computers & chemical engineering*, *27*(3), 327-346.

Virdhagriswaran, S., & Dakin, G. (2006, August). Camouflaged fraud detection in domains with complex relationships. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 941-947). ACM.

Wickham, H., & Stryjewski, L. (2011). 40 years of boxplots. *Am. Statistician*.

Wit, R.C. (2016). *Data-driven audit with anomaly detection algorithms an explorative study about the application of unsupervised machine learning to detect exceptions in transaction level audit data.* (non-publishes Master Thesis). Eindhoven University of Technology, Netherlands.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wong, S., & Venkatraman, S. (2015). Financial accounting fraud detection using business intelligence. *Asian Economic and Financial Review*, *5*(11), 1187.

Wu, G., & Chang, E. Y. (2003, August). Class-boundary alignment for imbalanced dataset learning. In *ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC* (pp. 49-56).

Xue, Z., Shang, Y., & Feng, A. (2010). Semi-supervised outlier detection based on fuzzy rough C-means clustering. *Mathematics and Computers in simulation*, *80*(9), 1911-1921.

Yuting, F. (2014). Analyzing European National Accounts Data for Detection of anomalous observation.

Zhang, D., & Lee, W. S. (2005, September). A simple probabilistic approach to learning from positive and unlabeled examples. In *Proceedings of the 5th Annual UK Workshop on Computational Intelligence (UKCI)* (pp. 83-87).

Zhang, J. (2013). Advancements of outlier detection: A survey. *ICST Transactions on Scalable Information Systems*, *13*(1), 1-26

Zhang, T., Ramakrishnan, R., & Livny, M. (1996, June). BIRCH: an efficient data clustering method for very large databases. In *ACM Sigmod Record* (Vol. 25, No. 2, pp. 103-114). ACM..

Zhang, Y., Meratnia, N., & Havinga, P. (2007). A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets. *Rap. tech., Centre for Telematics and Information Technology University of Twente*.

Annex

Annex A

Table 17: Frequency table for the number of observations and companies in the data set, after cleaning process.

Nb Obs:	Nb Companies:	Total Obs.	Freq.Rel (companies)	Freq.Rel (obs)	Freq.Rel.Ac (companies)	Freq.Rel.Ac (Obs)
1	219	219	2,30%	0,20%	2,30%	0,20%
2	282	564	2,96%	0,51%	5,27%	0,70%
3	487	1461	5,12%	1,31%	10,39%	2,01%
4	2930	11720	30,80%	10,51%	41,19%	12,52%
5	63	315	0,66%	0,28%	41,85%	12,80%
6	95	570	1,00%	0,51%	42,85%	13,31%
7	245	1715	2,58%	1,54%	45,42%	14,85%
8	1136	9088	11,94%	8,15%	57,36%	23,00%
9	49	441	0,52%	0,40%	57,88%	23,40%
10	54	540	0,57%	0,48%	58,45%	23,88%
11	134	1474	1,41%	1,32%	59,85%	25,20%
12	646	7752	6,79%	6,95%	66,65%	32,15%
13	25	325	0,26%	0,29%	66,91%	32,45%
14	50	700	0,53%	0,63%	67,43%	33,07%
15	128	1920	1,35%	1,72%	68,78%	34,80%
16	499	7984	5,25%	7,16%	74,03%	41,95%
17	19	323	0,20%	0,29%	74,22%	42,24%
18	34	612	0,36%	0,55%	74,58%	42,79%
19	106	2014	1,11%	1,81%	75,70%	44,60%
20	416	8320	4,37%	7,46%	80,07%	52,06%
21	24	504	0,25%	0,45%	80,32%	52,51%
22	43	946	0,45%	0,85%	80,77%	53,36%
23	105	2415	1,10%	2,17%	81,88%	55,52%
24	395	9480	4,15%	8,50%	86,03%	64,03%
25	28	700	0,29%	0,63%	86,32%	64,65%
26	32	832	0,34%	0,75%	86,66%	65,40%
27	85	2295	0,89%	2,06%	87,55%	67,46%
28	298	8344	3,13%	7,48%	90,69%	74,94%
29	38	1102	0,40%	0,99%	91,09%	75,93%
30	56	1680	0,59%	1,51%	91,67%	77,43%
31	178	5518	1,87%	4,95%	93,55%	82,38%
32	614	19648	6,45%	17,62%	100,00%	100,00%
Total	**9513**	**111521**	**100,00%**	**100,00%**	**100,00%**	**100,00%**

Annex B

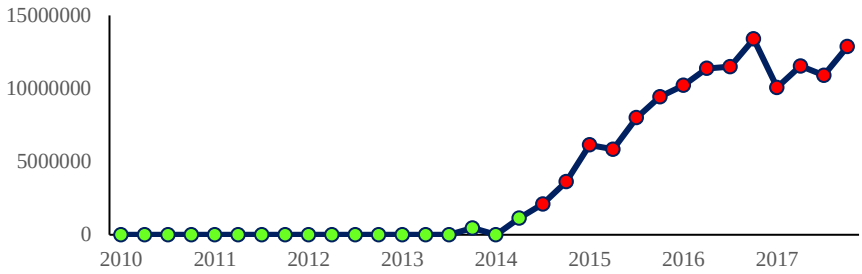

Figure 25 – Z-score method for IDcompany nº1200058152 and attribute B82.

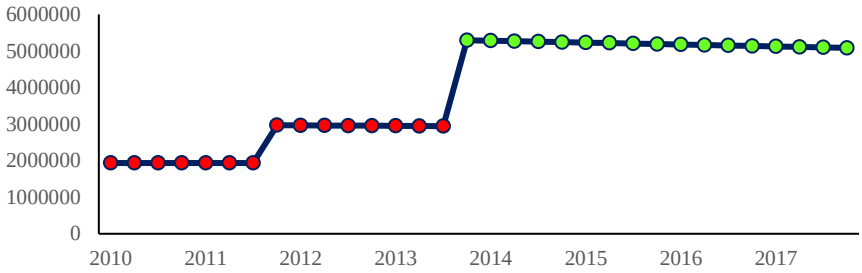

Figure 26 - Z-score method for IDcompany nº 1200058152 and attribute B82.

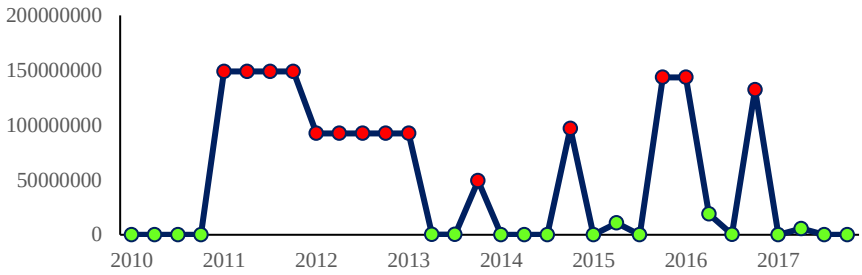

Figure 27 - Z-score method for IDcompany nº 1200272236 and attribute B82.

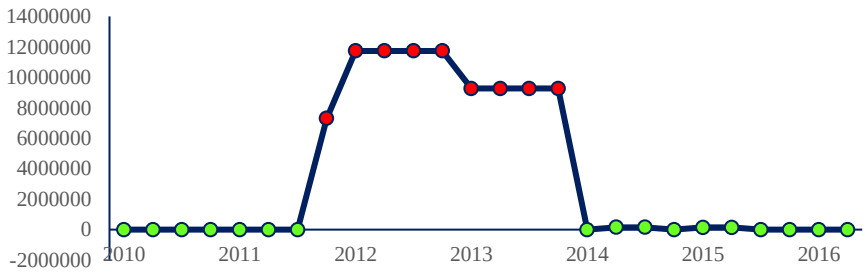

Figure 28 - Z-score method for IDcompany nº 1200272238 and attribute B82.

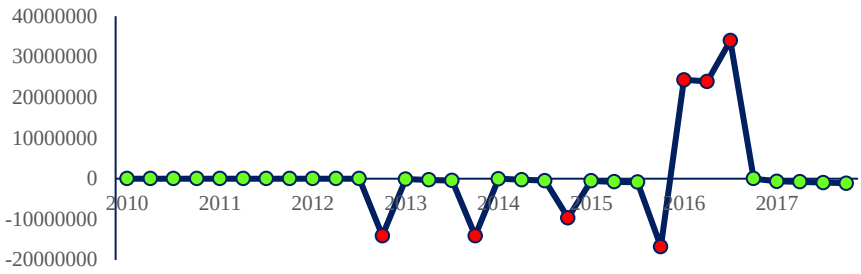

Figure 29 - Z-score method for IDcompany nº 1200650877 and attribute B82.

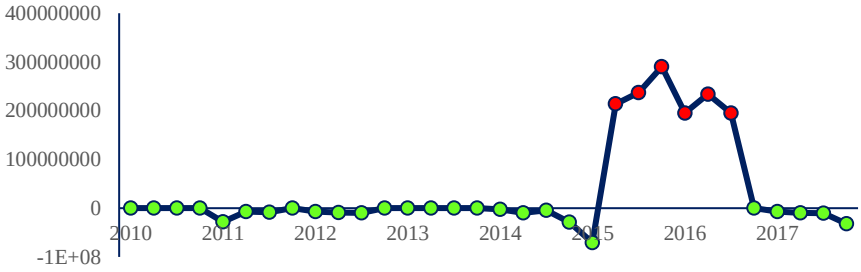

Figure 30 - Z-score method for IDcompany nº 1000019230 and attribute B82.

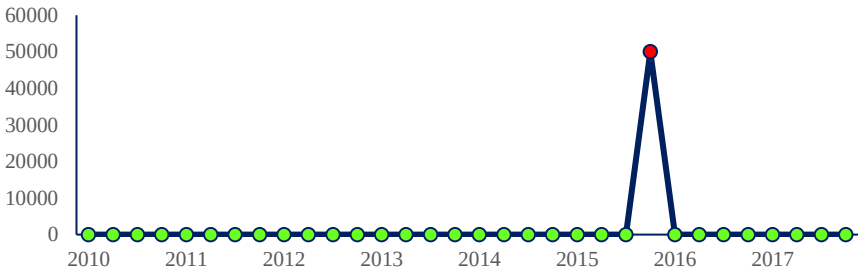

Figure 31 - Z-score method for IDcompany nº 1000049491 and attribute B82.

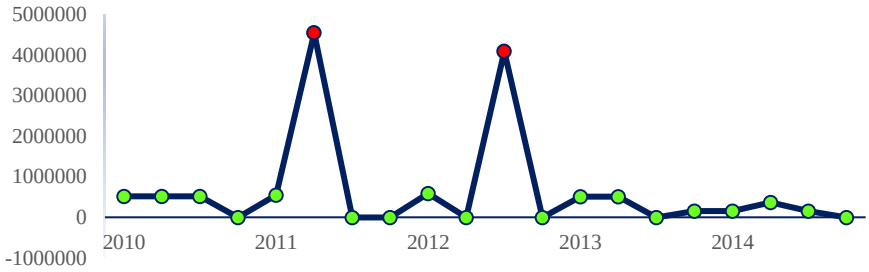

Figure 32 - Z-score method for IDcompany nº 1103055976 and attribute B82.

Annex C

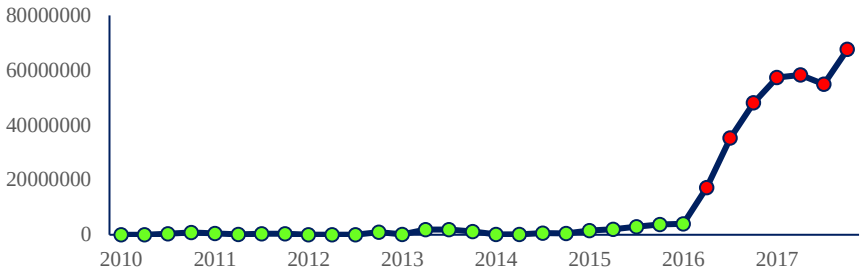

Figure 33 - Boxplot method for IDcompany nº 1220001527 and attribute B15.

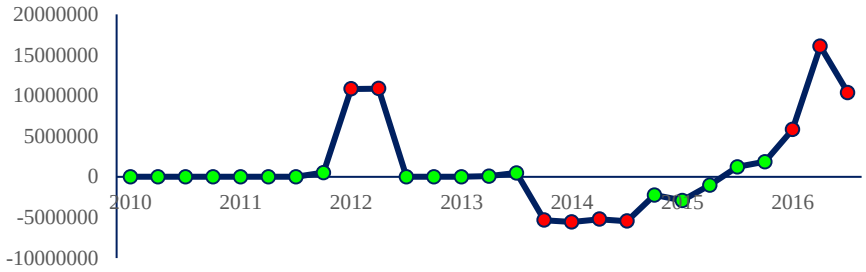

Figure 34 - Boxplot method for IDcompany nº 1000447030 and attribute B82.

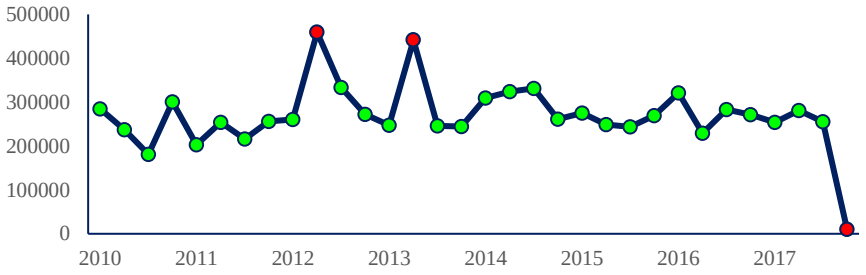

Figure 35 - Boxplot method for IDcompany nº 1000019802 and attribute B25.

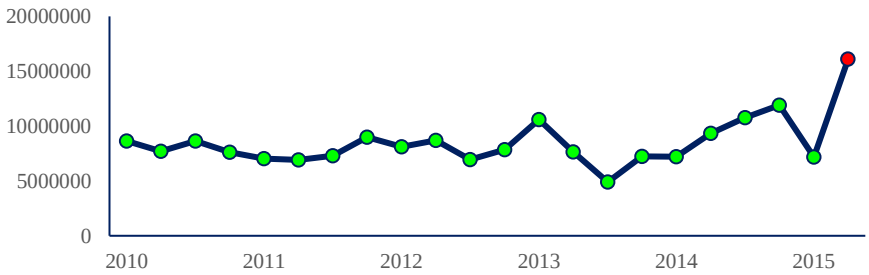

Figure 36 - Boxplot method for IDcompany nº 1000008889 and attribute B25.

Annex D

Table 18: Highest sequence of outliers for the outliers detected in the attributes with z-score for companies with less than 20 observations.

Nb Out	1	2	3	4	5	6	7	8	9	Total
1	8929									8929
2	1124	1392								2516
3	210	488	1084							1782
4	33	124	183	520						860
5	6	52	92	127	207					484
6		14	38	38	12	48				150
7		3	9	36	13	15	80			156
8			4	5	2	2	6	11		30
9			2	1	1	2	1	3	2	12
Total	10302	2073	1412	727	235	67	87	14	2	14919

Table 19: Highest sequence of outliers for the outliers detected in the attributes with z-score for companies with at least 20 observations.

Nb Out	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
1	2959															2959
2	1034	1057														2091
3	350	540	977													1867
4	123	297	335	933												1688
5	41	108	212	254	395											1010
6	19	52	149	140	95	215										670
7	5	31	79	132	69	51	240									607
8	2	17	43	69	44	23	55	162								415
9	1	16	36	39	45	30	38	41	136							382
10		4	15	26	35	25	39	28	22	84						278
11		2	12	13	21	21	27	34	25	13	107					275
12		1	8	6	10	12	21	14	12	13	12	54				163
13			5	6	7	10	19	8	17	16	16	9	67			180
14			1	7	4	1	10	10	6	6	6	9	13	46		119
15			2	2	1	1	9	5	5	3	4	7	5	7	48	99
Total	4534	2125	1874	1627	726	389	458	302	223	135	145	79	85	53	48	12803

Table 20: Highest sequence of outliers for the outliers detected in the attributes with boxplot for companies with at least 20 observations.

Nb Out	1	2	3	4	5	6	7	8	9	Total
1	2944									2944
2	734	1027								1761
3	313	406	832							1551
4	75	112	165	297						649
5	17	35	43	38	86					219
6	7	9	37	16	12	34				115
7	2	4	3	9	5	9	21			53
8			1	3	2		2	2		10
9				2	1		2	2	2	9
10				1						1
11								1		1
Total	4092	1593	1081	366	106	43	25	5	2	7313

Annex E

Table 21: Number of outliers detected per attribute for each CAE with z-score method for companies with less than 20 observations, in relative frequencies.

CAE	B05	B10	B15	B25	B30	B41	B51	B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	Total
1	1,80%	2,85%	4,12%	3,91%	2,33%	9,62%	8,03%	4,02%	2,75%	0,00%	4,02%	0,11%	4,97%	2,96%	1,06%	2,75%	2,75%	0,21%	3,24%
2	2,57%	2,33%	5,94%	4,72%	5,33%	7,47%	7,47%	4,10%	3,43%	0,92%	2,82%	0,55%	4,53%	2,57%	1,47%	2,39%	2,82%	0,00%	3,41%
3	2,15%	3,20%	3,72%	2,72%	4,61%	5,29%	5,87%	5,29%	3,93%	0,31%	2,46%	0,10%	2,62%	3,46%	0,84%	3,88%	2,36%	1,15%	3,00%
4	1,36%	1,97%	5,00%	1,74%	4,62%	6,07%	5,91%	3,26%	3,03%	0,61%	2,96%	0,15%	5,53%	4,85%	1,21%	3,11%	3,64%	0,15%	3,07%
5	0,00%	2,08%	0,00%	16,67%	14,58%	4,17%	8,33%	2,08%	0,00%	0,00%	2,08%	0,00%	2,08%	0,00%	2,08%	2,08%	2,08%	0,00%	3,24%
6	1,04%	3,59%	6,52%	3,31%	5,67%	7,93%	4,63%	5,10%	4,34%	0,00%	2,46%	0,57%	3,78%	4,34%	0,57%	4,53%	2,74%	0,38%	3,42%
7	1,87%	2,45%	4,07%	3,29%	3,94%	7,75%	4,91%	4,07%	4,65%	0,39%	2,52%	0,45%	4,78%	2,58%	0,84%	3,49%	3,29%	0,26%	3,09%
8	2,48%	2,21%	4,75%	2,43%	6,29%	5,69%	5,69%	2,26%	1,99%	0,28%	2,15%	0,44%	3,48%	4,42%	1,88%	2,87%	5,02%	0,61%	3,05%
9	1,30%	4,00%	5,42%	3,89%	5,89%	7,18%	4,48%	4,24%	2,71%	0,12%	2,12%	0,59%	5,06%	1,30%	0,00%	5,54%	1,53%	0,00%	3,08%
10	2,69%	3,15%	5,57%	3,90%	8,07%	8,16%	6,22%	3,90%	2,41%	0,56%	3,90%	0,37%	4,45%	6,31%	0,19%	1,48%	2,41%	0,37%	3,56%
11	1,66%	4,50%	5,67%	3,62%	5,38%	6,74%	6,84%	5,18%	4,11%	0,00%	2,54%	0,39%	2,74%	3,62%	1,08%	2,93%	4,01%	0,00%	3,39%
12	2,28%	3,09%	5,51%	3,63%	4,97%	7,06%	5,85%	5,44%	5,98%	0,07%	2,76%	0,13%	1,95%	2,76%	0,47%	3,23%	2,42%	0,00%	3,20%
13	1,99%	6,06%	1,16%	10,21%	4,15%	8,96%	6,97%	2,74%	2,07%	0,08%	2,49%	0,75%	4,90%	2,07%	0,00%	2,41%	4,56%	0,25%	3,43%
14	0,69%	2,30%	3,22%	2,42%	4,72%	7,13%	7,48%	5,06%	5,87%	0,00%	3,68%	0,92%	6,56%	0,92%	0,69%	3,57%	2,07%	0,00%	3,18%
18	2,20%	4,30%	2,59%	3,22%	3,13%	6,35%	6,45%	2,93%	3,22%	0,54%	1,76%	0,29%	4,98%	3,32%	1,17%	3,03%	1,95%	0,10%	2,86%
19	2,28%	4,57%	5,86%	3,89%	4,75%	7,34%	4,50%	5,43%	4,50%	0,31%	2,78%	0,62%	5,86%	4,38%	0,12%	5,18%	5,55%	0,19%	3,78%
20	2,08%	3,23%	5,54%	6,70%	3,00%	6,93%	5,08%	3,23%	2,31%	0,69%	4,16%	0,00%	5,54%	2,08%	0,23%	1,39%	0,92%	0,00%	2,95%
21	1,62%	6,87%	2,20%	6,24%	3,19%	6,73%	7,90%	2,60%	3,01%	0,04%	3,59%	0,31%	4,67%	2,38%	2,56%	4,98%	0,99%	0,22%	3,34%
22	2,12%	4,15%	3,59%	5,06%	4,97%	7,05%	6,57%	4,97%	3,80%	0,17%	2,38%	0,69%	5,23%	3,93%	0,65%	3,11%	2,29%	0,09%	3,38%
23	1,79%	3,73%	3,94%	5,78%	4,71%	7,62%	5,27%	4,14%	3,94%	0,10%	2,20%	0,46%	4,96%	2,51%	0,51%	3,38%	1,23%	0,41%	3,15%
24	2,14%	5,29%	1,80%	5,51%	1,24%	6,64%	3,94%	3,37%	1,91%	0,00%	2,02%	0,11%	3,71%	1,12%	0,45%	3,82%	0,11%	0,00%	2,40%
25	2,01%	5,10%	1,17%	4,01%	2,26%	8,44%	5,94%	3,01%	3,26%	0,59%	2,59%	0,50%	3,60%	1,92%	0,33%	1,84%	1,59%	0,08%	2,68%
26	1,39%	5,89%	4,68%	5,11%	3,90%	7,10%	5,28%	4,59%	5,19%	0,00%	2,60%	0,35%	5,71%	1,99%	0,69%	4,42%	2,42%	0,17%	3,42%
27	2,80%	6,36%	2,71%	5,26%	2,21%	6,79%	9,25%	5,51%	4,33%	1,95%	4,33%	2,80%	6,36%	4,92%	0,08%	2,80%	2,80%	0,42%	3,98%
41	1,83%	4,16%	3,52%	5,06%	2,78%	7,09%	7,73%	4,07%	3,92%	0,00%	4,16%	0,35%	5,06%	4,21%	2,33%	5,21%	1,93%	0,25%	3,54%
42	1,38%	6,22%	2,53%	4,92%	3,92%	6,30%	7,83%	3,76%	3,46%	0,46%	2,38%	0,23%	4,92%	3,53%	1,77%	3,53%	1,08%	0,38%	3,26%
43	2,54%	2,54%	4,42%	4,29%	4,88%	5,20%	5,07%	3,45%	2,08%	0,00%	2,28%	0,26%	5,27%	2,93%	1,30%	3,51%	1,17%	0,00%	2,84%
51	1,31%	2,43%	4,42%	3,24%	4,30%	5,79%	5,79%	4,86%	4,30%	0,93%	3,49%	0,06%	4,55%	1,99%	0,44%	3,86%	2,55%	0,37%	3,04%
52	1,60%	2,90%	6,05%	4,02%	5,70%	7,30%	5,94%	4,15%	4,13%	0,40%	1,92%	0,24%	4,79%	3,12%	1,17%	2,74%	3,33%	0,61%	3,34%
53	1,14%	4,45%	3,27%	4,21%	3,51%	6,21%	7,39%	2,94%	2,86%	0,41%	1,59%	0,04%	3,88%	2,00%	1,22%	3,14%	2,16%	0,20%	2,81%
60	2,78%	2,45%	3,41%	3,65%	4,65%	6,81%	6,53%	4,08%	4,61%	0,58%	2,83%	0,82%	4,27%	2,69%	0,77%	4,08%	1,54%	0,19%	3,15%
61	1,28%	6,06%	0,85%	7,12%	1,49%	9,88%	8,08%	4,46%	3,29%	0,11%	2,55%	0,11%	6,59%	4,68%	0,11%	5,74%	3,40%	0,64%	3,69%
Total	1,82%	3,89%	3,85%	4,80%	4,53%	7,03%	6,35%	4,01%	3,48%	0,33%	2,77%	0,43%	4,61%	3,00%	0,88%	3,44%	2,46%	0,24%	3,22%

Table 22: Number of outliers detected per attribute for each CAE with z-score method for companies at least 20 observations, in relative frequencies.

CAE	B05	B10	B15	B25	B30	B41	B51	B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	Total
1	3,83%	2,62%	2,42%	2,11%	2,11%	9,97%	6,75%	5,64%	8,26%	0,60%	4,93%	3,32%	6,85%	8,76%	1,81%	2,32%	3,52%	1,61%	4,30%
2	3,64%	2,02%	4,52%	3,84%	6,15%	7,45%	8,36%	4,65%	6,47%	2,21%	3,93%	1,17%	9,01%	7,25%	2,05%	2,11%	5,92%	1,04%	4,54%
3	5,97%	2,28%	3,10%	2,20%	4,48%	5,86%	6,98%	5,11%	4,85%	2,01%	3,40%	0,60%	6,75%	7,61%	1,83%	4,78%	6,01%	0,97%	4,15%
4	4,94%	2,86%	6,00%	2,08%	5,08%	8,68%	5,40%	4,62%	4,62%	2,59%	3,97%	2,17%	8,55%	9,33%	2,36%	3,88%	5,73%	0,55%	4,63%
5	2,60%	1,30%	9,09%	3,90%	10,39%	6,49%	10,39%	9,09%	7,79%	0,00%	1,30%	0,00%	7,79%	0,00%	0,00%	0,00%	0,00%	0,00%	3,90%
6	5,69%	3,08%	7,93%	1,98%	5,06%	7,17%	7,97%	5,31%	9,15%	1,43%	6,54%	1,14%	8,44%	8,81%	1,94%	1,90%	8,27%	0,42%	5,12%
7	2,50%	1,84%	3,11%	1,65%	3,34%	6,26%	5,49%	6,84%	7,76%	1,88%	3,92%	1,04%	8,30%	9,76%	1,84%	3,27%	7,57%	1,19%	4,31%
8	3,51%	3,00%	3,28%	2,49%	5,77%	8,73%	5,17%	5,45%	5,73%	2,08%	3,46%	0,55%	8,04%	7,44%	1,25%	3,79%	8,13%	0,88%	4,38%
9	1,73%	4,18%	4,74%	2,51%	3,63%	8,42%	6,02%	7,81%	5,86%	1,78%	6,25%	1,34%	8,92%	8,64%	1,28%	5,41%	6,25%	0,39%	4,73%
10	3,20%	1,92%	2,77%	2,41%	4,47%	5,89%	7,81%	6,25%	7,17%	1,14%	4,26%	0,92%	4,55%	5,47%	2,41%	2,41%	6,25%	0,00%	3,85%
11	2,98%	1,85%	3,61%	1,89%	4,37%	7,12%	7,48%	6,67%	7,17%	1,76%	4,42%	1,49%	5,09%	3,88%	0,81%	2,21%	4,91%	1,35%	3,84%
12	5,54%	3,67%	4,34%	3,34%	5,41%	6,41%	5,61%	5,61%	8,02%	1,47%	2,94%	0,80%	5,61%	6,35%	1,60%	2,27%	4,74%	0,13%	4,10%
13	4,03%	4,51%	4,18%	9,44%	6,31%	8,02%	6,69%	4,89%	3,37%	0,71%	8,92%	1,14%	8,83%	4,60%	1,23%	2,42%	6,98%	1,00%	4,85%
14	1,93%	2,89%	6,05%	5,61%	5,26%	8,95%	7,89%	2,81%	3,86%	0,26%	2,89%	0,18%	5,61%	6,23%	2,02%	1,93%	4,30%	0,44%	3,84%
18	2,63%	3,27%	3,13%	3,77%	6,90%	8,33%	7,76%	6,48%	7,05%	1,21%	4,77%	1,92%	7,33%	4,91%	1,85%	4,91%	4,84%	0,64%	4,54%
19	3,17%	3,36%	6,97%	4,42%	7,71%	7,15%	6,65%	9,33%	8,64%	1,06%	5,16%	1,06%	10,14%	9,89%	0,19%	6,90%	9,58%	0,25%	5,65%
20	6,97%	2,65%	10,74%	3,49%	7,11%	10,46%	8,79%	11,02%	9,34%	0,28%	8,93%	3,77%	10,04%	5,16%	1,81%	3,07%	8,93%	1,26%	6,32%
21	5,37%	9,72%	1,58%	8,21%	4,67%	8,84%	9,15%	4,55%	4,80%	0,00%	6,31%	2,21%	8,59%	5,74%	4,55%	6,19%	4,99%	0,51%	5,33%
22	4,68%	5,47%	7,56%	5,19%	7,78%	7,22%	7,33%	6,03%	5,41%	0,96%	5,30%	2,03%	10,04%	4,96%	0,85%	6,20%	4,91%	0,39%	5,13%
23	3,48%	2,46%	5,94%	6,42%	7,86%	6,60%	8,22%	6,42%	7,80%	1,56%	4,68%	2,76%	7,98%	7,20%	1,74%	4,08%	7,62%	1,02%	5,21%
24	6,12%	4,03%	3,62%	2,50%	3,20%	7,79%	6,40%	5,70%	6,54%	0,00%	5,42%	0,14%	7,37%	7,23%	1,67%	1,53%	5,15%	0,70%	4,17%
25	2,18%	3,51%	2,47%	3,51%	2,75%	9,78%	7,98%	5,60%	4,46%	5,32%	6,55%	1,99%	8,64%	9,69%	0,66%	2,47%	6,55%	0,19%	4,69%
26	1,67%	4,07%	4,65%	4,58%	4,36%	9,45%	5,67%	3,85%	5,01%	1,96%	4,14%	0,65%	7,34%	6,54%	3,63%	2,91%	2,54%	0,07%	4,06%
27	7,74%	10,28%	4,45%	11,76%	6,09%	10,44%	13,88%	7,00%	8,21%	5,67%	7,47%	8,37%	13,30%	3,82%	1,27%	7,90%	7,10%	2,33%	7,62%
41	2,11%	4,23%	3,92%	5,63%	6,44%	9,86%	11,97%	2,72%	4,53%	0,80%	3,92%	1,41%	5,73%	6,14%	5,03%	6,04%	3,62%	3,32%	4,86%
42	8,83%	4,38%	6,31%	5,64%	8,90%	5,64%	5,71%	4,75%	3,34%	2,82%	5,04%	4,97%	11,13%	8,23%	3,49%	3,64%	10,53%	3,19%	5,92%
43	3,61%	2,40%	4,70%	5,25%	7,43%	4,04%	4,92%	6,67%	4,81%	1,20%	2,19%	1,31%	7,98%	5,90%	3,72%	6,67%	6,56%	0,44%	4,43%
51	4,03%	3,43%	10,14%	3,73%	5,35%	10,14%	4,84%	6,40%	6,51%	1,46%	4,64%	1,77%	7,51%	7,41%	1,66%	7,82%	6,76%	0,25%	5,21%
52	4,00%	2,34%	7,06%	2,60%	5,78%	8,55%	7,56%	7,14%	7,43%	1,57%	4,49%	1,54%	7,24%	6,44%	1,72%	4,15%	6,19%	0,65%	4,80%
53	4,81%	6,49%	7,06%	3,87%	4,40%	9,58%	9,37%	6,33%	8,27%	1,36%	3,98%	1,62%	10,41%	9,73%	2,09%	1,36%	5,08%	0,99%	5,38%
60	3,14%	5,90%	3,18%	4,00%	5,08%	7,14%	6,28%	7,47%	5,57%	1,20%	3,44%	1,05%	8,56%	5,68%	0,37%	5,61%	5,94%	0,49%	4,45%
61	2,30%	4,37%	3,28%	5,51%	4,94%	8,71%	7,50%	5,09%	5,28%	1,11%	3,09%	0,21%	5,34%	5,43%	1,79%	4,90%	7,20%	0,96%	4,28%
Total	4,03%	3,76%	5,06%	4,24%	5,58%	7,97%	7,44%	6,04%	6,35%	1,55%	4,71%	1,71%	8,03%	6,70%	1,89%	3,91%	6,02%	0,86%	4,75%

Table 23: Number of outliers detected per attribute for each CAE with boxplot method for companies at least 20 observations, in relative frequencies.

CAE	B05	B10	B15	B25	B30	B41	B51	B60	B65	B70	B76	B78	B82	C50	C60	C75	C80	C90	Total
1	1,11%	0,91%	1,01%	1,01%	1,41%	2,82%	0,91%	2,11%	2,42%	0,20%	1,31%	1,11%	2,72%	1,11%	0,81%	1,51%	0,91%	0,30%	1,31%
2	2,41%	0,42%	1,69%	1,14%	1,63%	1,89%	1,69%	1,07%	2,18%	0,68%	1,53%	0,81%	4,49%	1,89%	0,78%	0,81%	2,80%	0,91%	1,60%
3	2,95%	0,63%	0,71%	0,37%	1,42%	1,49%	1,23%	0,82%	1,46%	0,60%	1,34%	0,37%	4,14%	1,53%	1,08%	1,23%	1,19%	0,11%	1,26%
4	2,08%	0,32%	1,89%	0,74%	1,57%	1,94%	1,20%	0,74%	0,83%	0,65%	1,29%	0,83%	3,97%	2,17%	0,79%	0,60%	1,29%	0,23%	1,29%
5	0,00%	1,30%	2,60%	1,30%	2,60%	0,00%	5,19%	0,00%	0,00%	0,00%	1,30%	0,00%	7,79%	0,00%	0,00%	0,00%	0,00%	0,00%	1,23%
6	2,49%	1,05%	1,73%	0,30%	1,05%	1,56%	1,43%	1,48%	1,64%	0,38%	2,02%	0,51%	2,61%	1,52%	0,72%	0,72%	1,73%	0,51%	1,30%
7	1,92%	0,31%	1,42%	0,85%	0,85%	1,34%	0,69%	1,23%	2,38%	0,65%	1,11%	0,96%	3,84%	2,31%	0,69%	0,88%	1,96%	0,42%	1,32%
8	2,36%	1,20%	0,92%	0,51%	1,34%	1,71%	1,34%	1,66%	1,39%	0,28%	1,34%	0,60%	3,09%	1,29%	0,55%	0,92%	1,66%	0,14%	1,24%
9	1,06%	0,50%	1,06%	0,89%	1,28%	2,01%	1,23%	2,29%	2,18%	0,28%	1,56%	0,56%	4,07%	1,45%	0,50%	1,51%	2,96%	0,39%	1,43%
10	1,49%	0,99%	0,57%	0,57%	1,92%	2,20%	0,57%	1,35%	1,21%	0,64%	1,49%	0,50%	2,70%	1,42%	0,78%	1,42%	1,35%	0,00%	1,18%
11	1,53%	0,81%	0,99%	0,36%	1,31%	1,31%	0,99%	1,67%	2,30%	0,45%	2,30%	0,90%	2,34%	1,08%	0,45%	0,54%	1,44%	0,27%	1,17%
12	2,40%	1,47%	1,14%	0,53%	1,07%	1,94%	1,54%	0,60%	1,47%	0,80%	1,14%	0,67%	2,47%	1,00%	0,94%	1,14%	2,20%	0,13%	1,26%
13	1,95%	1,23%	1,38%	2,75%	2,18%	0,95%	1,61%	0,95%	1,28%	0,52%	0,85%	0,43%	4,22%	2,52%	0,00%	1,00%	2,04%	0,85%	1,48%
14	1,32%	0,61%	2,19%	0,70%	2,37%	1,58%	2,11%	1,75%	1,93%	0,26%	0,35%	0,18%	3,25%	1,58%	0,61%	1,93%	2,02%	0,44%	1,40%
18	2,14%	1,07%	2,21%	1,71%	1,99%	2,63%	0,93%	1,14%	2,06%	0,43%	2,63%	1,14%	2,99%	0,93%	0,71%	1,71%	1,00%	0,28%	1,54%
19	2,67%	1,00%	1,49%	0,62%	2,05%	1,24%	0,62%	2,61%	2,55%	0,37%	1,37%	0,75%	3,23%	1,80%	0,19%	0,81%	2,80%	0,00%	1,45%
20	2,65%	2,23%	2,37%	0,56%	1,53%	2,93%	2,65%	3,07%	1,39%	0,28%	1,95%	1,12%	5,72%	3,91%	0,42%	0,42%	1,67%	0,84%	1,98%
21	1,01%	1,33%	0,44%	2,78%	2,15%	1,39%	2,08%	1,52%	1,96%	0,00%	1,20%	1,14%	2,65%	1,45%	1,83%	2,27%	2,21%	0,25%	1,54%
22	1,47%	1,64%	2,31%	1,80%	3,05%	1,64%	1,80%	0,85%	1,41%	0,11%	1,18%	1,47%	4,12%	1,58%	0,17%	1,41%	1,86%	0,17%	1,56%
23	1,56%	0,66%	2,94%	1,20%	3,06%	1,38%	1,50%	1,50%	2,34%	0,36%	1,56%	1,26%	4,86%	1,92%	0,42%	0,66%	2,88%	0,66%	1,71%
24	0,70%	1,67%	1,39%	0,42%	1,11%	1,81%	0,97%	2,36%	2,09%	0,00%	0,97%	0,00%	2,36%	2,36%	0,97%	0,70%	1,81%	0,70%	1,24%
25	1,71%	1,14%	1,42%	1,04%	1,33%	1,33%	1,23%	1,61%	1,61%	0,66%	1,33%	0,66%	2,47%	1,04%	0,95%	0,95%	1,52%	0,19%	1,23%
26	1,16%	1,74%	2,40%	0,87%	2,18%	2,69%	1,24%	1,24%	0,94%	0,00%	1,16%	0,65%	3,56%	2,11%	0,65%	1,89%	0,80%	0,07%	1,41%
27	3,29%	3,13%	2,07%	4,56%	2,76%	2,86%	2,86%	1,75%	2,76%	0,85%	3,02%	2,97%	3,97%	0,58%	0,37%	0,74%	2,60%	0,95%	2,34%
41	0,80%	0,91%	1,51%	2,31%	2,11%	1,91%	1,91%	1,61%	1,51%	0,20%	1,41%	0,20%	3,12%	1,41%	1,11%	1,71%	1,81%	1,11%	1,48%
42	2,97%	1,04%	1,19%	1,63%	2,00%	1,63%	0,67%	1,26%	1,41%	0,30%	2,74%	2,30%	2,23%	1,41%	1,41%	0,74%	1,71%	2,52%	1,62%
43	1,42%	1,31%	1,53%	0,33%	1,09%	0,55%	1,09%	1,97%	1,53%	0,33%	1,97%	1,31%	1,97%	1,09%	1,53%	0,55%	1,09%	0,44%	1,17%
51	1,97%	1,01%	2,57%	1,16%	1,51%	1,51%	0,71%	2,27%	2,47%	0,86%	1,31%	0,61%	2,82%	1,46%	0,86%	2,07%	1,36%	0,25%	1,49%
52	2,25%	0,55%	2,01%	1,11%	1,77%	2,78%	1,61%	2,48%	2,74%	0,30%	1,89%	0,90%	3,27%	1,63%	0,47%	0,94%	2,08%	0,43%	1,62%
53	2,15%	3,35%	2,25%	1,10%	0,94%	2,30%	1,57%	1,88%	2,46%	0,37%	0,99%	1,15%	4,03%	2,51%	0,84%	0,73%	0,99%	0,26%	1,66%
60	1,68%	1,98%	1,53%	1,57%	1,83%	1,16%	0,67%	1,72%	1,91%	0,37%	1,05%	0,56%	2,24%	1,16%	0,49%	0,90%	1,57%	0,11%	1,25%
61	1,11%	1,30%	1,13%	1,72%	1,70%	1,81%	1,38%	1,87%	2,11%	0,34%	0,45%	0,13%	2,88%	1,34%	0,26%	1,00%	1,55%	0,64%	1,26%
Total	1,80%	1,21%	1,63%	1,20%	1,75%	1,76%	1,48%	1,58%	1,81%	0,39%	1,47%	0,84%	3,44%	1,58%	0,70%	1,08%	1,71%	0,46%	1,44%

Annex F

Table 24: Estratification for number the of observations per CAE and size, considering the two subsets of observations used in the practical implementation.

	Companies from 4 to 19 observations				**Companies from 20 to 32 observations**			
	Size				Size			
CAE	1	2	3	4	1	2	3	4
1	312	566	71	0	143	451	295	104
2	211	568	759	95	86	414	1299	1276
3	173	574	1019	143	48	512	1171	949
4	222	550	506	41	121	547	838	659
5	28	16	0	4	13	32	0	32
6	251	473	309	26	147	418	1091	715
7	269	624	632	23	84	326	1100	1093
8	214	787	730	80	79	400	1091	595
9	232	403	201	13	65	344	677	707
10	302	471	280	25	60	419	680	249
11	145	427	396	55	78	298	845	997
12	292	633	540	23	80	393	650	374
13	448	500	235	22	265	505	713	624
14	259	416	152	42	127	401	478	134
18	280	738	837	193	199	220	446	540
19	341	550	566	164	77	223	613	695
20	166	165	55	47	92	142	285	198
21	1374	665	189	0	780	399	329	76
22	915	829	415	154	359	353	719	342
23	528	715	477	235	104	390	389	784
24	248	353	280	8	102	200	264	153
25	522	367	247	60	191	218	204	440
26	455	414	236	50	355	347	503	171
27	931	222	22	4	837	748	270	32
41	772	727	398	120	203	311	297	183
42	330	466	407	99	65	159	456	668
43	433	649	381	75	170	207	287	251
51	408	528	612	57	135	296	917	635
52	865	1269	1468	153	346	961	3190	2075
53	903	874	455	217	104	200	323	1284
60	537	824	514	209	276	634	977	789
61	265	350	250	73	591	1563	2003	1147
Total	13631	17713	13639	2510	6382	13031	23400	18971

Annex G

Table 25: Comparison of the detected outliers for the different methods, for companies with at least 20 observations.

	LOF k=3	LOF k=19	DBSCAN k=4	DBSCAN K=20	Boxplot	Z-score	Out (%)
LOF k=3	**2296**	809	526	538	869	1468	3,72%
LOF k=19		**2390**	984	1123	1535	2080	3,87%
DBSCAN k=4			**1617**	1617	1480	1562	2,62%
DBSCAN K=20				**2118**	1906	2049	3,43%
Boxplot					**12458**	11485	20,16%
Z-score						**31271**	50,61%